Inferring Temporal Dynamics of Gene Expression from RNA Editing Profiles

James A. C. Bayne

Supervisors:

Prof. Mark Coles, Prof. Eamonn Gaffney, Dr. Sam Rodriques

St. Hilda's College

University of Oxford

Submitted in partial fulfilment of the requirements for the degree of Doctorate of Philosophy

April 2025

Memorandum

The research, results and their interpretation in this thesis are my own, unless explicitly stated otherwise.

This research was funded with support from the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2168), the UK Medical Research Council (CC2168), and the Wellcome Trust (CC2168). It was also funded by the Impetus Grant programme as well as with support and from Eric and Wendy Schmidt.

My studentship was provided by the EPSRC via the Sustainable Approaches to Biomedical Science, Responsible and Reproducible Research Centre for Doctoral Training (EP/S024093/1) as well as from the Francis Crick Institute.

I used the data from the following experiments performed by Ali Ghareeb at the Crick: all of the calibration experiments (listed in Table 2.1: the hiPSC-derived cultures were with made the help of James Evans and Minee Lee Choi from the Ghandi lab at the Crick (fibroblasts from healthy donors were reprogrammed with approval from the London-Hamstead Research Ethics Committee and the University College London, Great Ormond Street Institute of Child Health and Great Ormond Street Hospital Joint Research Office), the response of HEK cells to heat shock (detailed in Sections 3.5, 4.3), the human intestinal organoid culture (Section 4.5, under Research Ethics Committee references 04-Q0508-79 and 18/EE/0150) and human organotypic brain slices (Section 4.5, under National Research Ethics approval reference: 21/SC/0111).

The monocyte isolation was performed by Gabrielle Chappell at The Kennedy Institute of Rheumatology at the University of Oxford. The monocyte stimulation with lipopolysaccharide was performed by Ali Ghareeb. Monocytes were isolated from

the blood of a healthy male via the UK National Health Service blood bank.

The HEK single cell sequencing was performed by Aaron Wagen and Ali Ghareeb.

Thank you to the following people for their contributions: Prof. Eamonn Gaffney for his help formalising the mathematical model of editing, Rory Maizels for providing the empty Tet-On plasmid used in Figures 4.1 and 4.2, Danny Lang at the Scientific Computing STP at the Crick for help with high-performance computing, Neelam Mehta, Danson Loi, Eleanor Calcutt and Adam Cribbs from the Cribbs lab at the Botnar Institute for Musculoskeletal Sciences at the University of Oxford for help with Oxford Nanopore Sequencing and for letting us use their Promethion24 sequencer. To the Long Read Sequencing Facility at University College London for providing access to their Pacific Biosciences long-read sequencer and to George Young for supplying Nextflow pipelines for processing of the raw sequencing data.

This thesis has only been submitted to the University of Oxford for the degree of Doctor of Philosophy in Cellular and Molecular Medicine and not for any other degree at the same or any other university.

This thesis contains approximately 45,000 words.

Acknowledgements

It is fitting that the first reference of my thesis is of Sir Francis Crick's 1970 paper on the central dogma, given that much of the research that I describe here was done during my time at an institute bearing his name. I feel very fortunate to have undertaken my DPhil research both at Oxford and at the Crick, getting the best of both worlds (as well as a significant amount of time on the British trains, which cannot be described as a 'best world'). That I was able to do this is, and indeed everything else to do with my DPhil, is thanks to my supervisors: Mark, Eamonn and Sam. Mark and Eamonn - you took me on as a SABS rotation student in the midst of a pandemic in the Spring of 2021, before we even had a chance to meet in person and have been such incredibly supportive supervisors ever since. You gave me ample space to explore my own research avenues and those at the Crick, whilst always giving me feedback at our weekly meetings and further help when needed. You helped me navigate the changes that happened with the Crick lab in the last year and get me to submission - thank you!

Sam, you were the reason I found my way to the Crick and I'm very grateful for the effort you put in to make that happen. I joined your lab without knowing what project I would be working on but with an intuition that it was the right place to be - 4 years on I believe my hunch was right. I got to see the full life cycle of a lab and met so many incredible people along the way. For a project that required tremendous amounts of sequencing, I feel very privileged to have been in a lab where each year we were told that we weren't spending enough money!

For my research itself, I want to thank Ali and Aaron, without whom I would have no data! I also want to thank George, Laura C, Laura R, Adam, Neelam, Gabrielle and David for their contributions during the project.

I of course owe a huge debt of thanks to the support network around me, especially

Mum and Dad, to whom I can't do justice here. To my sisters: Ally, Bina and Tash and all my relatives. One of the benefits of commuting to London was being able to see so many of my friends from my undergrad at Imperial (I'm not going to list everyone here but in particular: Alfonso, Amy, Michal, Lydia (a bit further away), Kevin, Alice and Marti - since many of you let me crash on your sofas). To Freddie for being such an amazing friend for 14 years despite now being stuck on the other side of the world.

To Victor, for hiring 18-year-old me as an intern at an 8 person CRISPR start-up in 2015: one of the big crossroads in my life. To Peter, the self-described 'angriest man in the UK' for being a great friend and mentor for the last 6 years - bringing me into 2 companies and demystifying Oxford for me. I hope I more or less returned the favour with the intro to Adam!

And then of course to everyone at Nucleate. To Pia, Josh and Tony at Petri and Michael, Oliver and Souf for bringing me in. To the founding UK group: Fede, Miro, Georg, Isabel, Abhi, Steph, Alissa, Alon, Alya, Olive and Amit. To the amazing people who are running it as I write this: Fede, Maja and Alya - there's so much I could write about how amazing each of you is but I'll save the ink. To the other Directors past and present. To Juan and Jason. The list goes on. I have loved working with all of you and I am incredibly proud of what we have built - I truly believe it is something special.

To everyone I haven't mentioned - thank you for supporting me, in whatever way, big or small.

"Show me your friends and I'll show you your future"

- Michael Retchin, 2023

Abstract

Human gene expression involves the continuous synthesis and decay of mRNA. The expressed genes of a cell at any given point in time can be measured by RNA sequencing (RNA-seq). However, RNA-seq only provides an instantaneous snapshot of the state of the cell and dedicated experimental protocols are required to measure the synthesis and decay kinetics. In this thesis, I develop new computational methods of Endogenous RNA Age (ERA) that extract temporal information from RNA-seq data. By measuring the underlying rate of A-to-I editing at all editing sites in the transcriptome, I show that these rates are robust, and that A-to-I edits encode the ages of human transcripts. The mean age of a population of transcripts can be estimated from short-read RNA-seq data, whilst long-read data can be used to further infer the ages of individual transcripts. Crucially, in contrast to current metabolic labelling methods, ERA requires no dedicated experimental protocols provided that the editing rates are known. I show that transcripts ages provide information on the kinetics of mRNA decay and - when combined with abundance measurements - on synthesis rates and past changes to transcription. This thesis presents the first characterisation of endogenous A-to-I editing rates in human cells and new methods for measuring mRNA kinetics from bulk and single-cell RNA sequencing data.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

"There are more things in Heaven and Earth, Horatio,
than are dreamt of in your philosophy."

– William Shakespeare, *Hamlet, Act 1, Scene 5*

1.1 Preface

During an hour-long lecture at University College London in 1957, Sir Francis Crick introduced the central dogma of molecular biology. Frequently misrepresented, misquoted and misunderstood as 'DNA makes RNA makes Protein' - the schematic (as eventually published 13 years later (Crick, 1970)) features not two information exchanges (DNA to RNA, RNA to protein), but six. Predating the discovery of messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA) (not to mention the plethora of other RNA species), the position of RNA at the centre of Crick's schematic proved prescient of its role as the key information mediator between nucleic and amino acids. As transient molecules, understanding the lifespan of mRNAs is crucial for deciphering gene expression dynamics. The typical human mRNA goes through a journey of transcription, maturation, translation and degradation - although many routes can be taken by any given molecule.

The research described in this thesis created methods to probe the lifecycle of mRNAs. This is achieved by modelling the degree of editing on transcripts from RNA sequencing data and demonstrated in several human cell lines and types. This chapter reviews several key concepts that underpin the methods including mRNA turnover, RNA editing by ADAR proteins and advances in RNA sequencing tech-

nologies (short-read, long-read, single-cell).

1.2 The lifecycle of a human mRNA

The life of an mRNA begins with transcription by RNA polymerase II, which can take anywhere from several minutes to over an hour depending on gene length and polymerase speed (typically 1-4 kilobases per minute) (Jonkers and Lis, 2015). A typical 26 kb gene might be transcribed in under 10 minutes, whereas a very large gene of 150 kb could require several hours of transcription (Piovesan et al., 2019). During this time, processing of the nascent transcript begins, with events such as 5' capping, splicing, adenosine-to-inosine (A-to-I) editing and RNA binding protein (RBP) recruitment (Figure 1.1). Most introns are spliced out within minutes, although splicing also occurs post-transcriptionally (Alpert et al., 2016). After transcription termination and any remaining processing, the mature mRNA and associated RBPs - collectively the messenger ribonucleoprotein (mRNP) - is exported out of the nucleus via the nuclear pore complex, typically on the order of 5-40 minutes after transcription termination, although some recent studies using metabolic labelling rather than imaging suggest longer time scales (Ben-Ari et al., 2010; Li et al., 2018; Müller et al., 2024). Following export into the cytosol, mRNA can then be used for protein translation, can undergo further modification and is ultimately degraded (Figure 1.1).

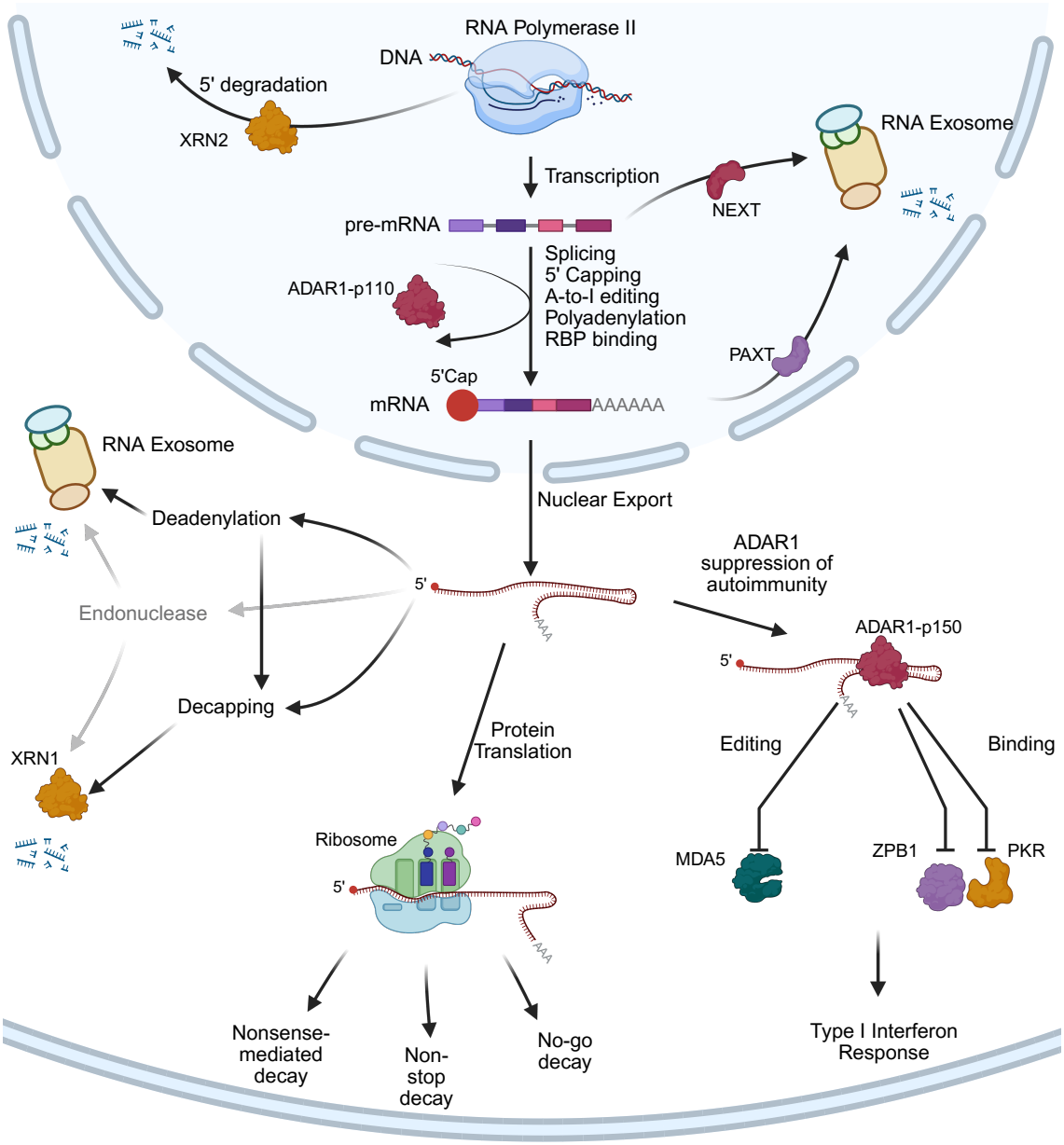

Figure 1.1: **Major pathways of human mRNA metabolism and A-to-I editing.** pre-mRNA is produced during transcription by RNA Polymerase II. Transcripts can be co-transcriptionally degraded by the 5'-3' exoribonuclease XRN2. During transcription, pre-mRNA receives a 5' 7-methylguanosine (m^7G) cap to protect against XRN2, can have introns spliced out by the spliceosome, is subject to A-to-I editing by ADAR1-p110, is polyadenylated and bound by various RNA binding proteins (RBPs). 5'-capped pre-mRNA/mRNA can be recruited to the nuclear RNA exosome for degradation by either NEXT (transcripts with no poly(A)-tail) or PAXT (transcripts with a poly(A)-tail). Once ready for nuclear export, the mature mRNA is transported to the cytosol for translation. In the cytosol, the mRNA can undergo a plethora of events. The main interactions with ADAR1-p150 are shown: editing of dsRNA regions suppresses MDA5-mediated type I interferon response. Binding of ADAR1-p150 to dsRNA regions suppresses both PKR and ZBP1 activation of interferon response, apoptosis and necroptosis. The major pathways of cytosolic mRNA degradation are shown: 5'→3' by XRN1 or 3'→5' by the cytosolic RNA exosome. Quality control during translation involves three main pathways: nonsense-mediated, non-stop and no-go decay. (Created with biorender.com)

The cytosolic mRNA degradation pathways are complex and intertwined but can be generally divided into three types based on ribonuclease activity: 5'→3' exoribonuclease (e.g. XRN1), 3'→5' exoribonuclease (e.g. the exosome complex) or by *endo*ribonucleases (Figure 1.1). The triggers for these mechanisms can be grouped into quality control mechanisms (QC) or regular turnover - although they share overlaps. Regular turnover of mRNA typically beings with deadenylation of the 3' poly(A)-tail by PAN2/PAN3, CCR4-NOT or PARN (Zhang et al., 2015; Yi et al., 2018). This exposes the 3' end of the transcript to the exosome, which can digest the mRNA in the 3'→5' direction. Deadenylation can also precede digestion of the mRNA from the 5' end, which can only happen after removal of the 5'-cap (decapping). Decapping by DCP2/DCP3 exposes the 5' end of the mRNA to the 5'→3' exoribonuclease XRN1 and can occur either after deadenylation or independently (Schoenberg and Maquat, 2012; Dowdle and Lykke-Andersen, 2025).

QC mechanisms describe the degradation of transcripts that have faulty translation. The three main QC pathways are non-sense mediated decay (NMD), non-stop decay (NSD) and no-go decay (NGD) (D'Orazio et al., 2019; D'Orazio and Green, 2021; Kögel et al., 2024). NMD is the best characterised and degrades mRNAs with a premature termination codon (PTC). During a 'pioneer' round of translation, the ribosome normally displaces any exon junction complexes (EJCs) present on the mRNA, which can be deposited upstream of exon junctions after splicing (Hir et al., 2015). If a stop codon is encountered while EJCs remain bound further downstream (an indication that the stop is premature), NMD is triggered. The stalled ribosome together with the EJC and associated factors mark the mRNA for rapid decay by both XRN1 (5'→3') and the exosome-SKI238 complex (3'→5') (Kögel et al., 2024). NMD prevents production of truncated proteins from faulty mRNAs (e.g. those arising from nonsense mutations or mis-splicing). Non-stop decay, on the other hand, removes mRNAs lacking an in-frame stop codon (the ribosome runs into the poly(A) tail) and decay mostly proceeds via deadenylation and 3'→5' digestion by the exosome (Arribere and Fire, 2018). Finally, no-go decay is activated when a

ribosome stalls on a problematic mRNA sequence or structure and tends to recruit an endonuclease, which cuts the mRNA internally, resulting in unprotected ends which can degraded by exoribonucleases both in both 5'→3' and 3'→5' directions (D'Orazio et al., 2019). Collectively, these surveillance pathways constitute a diverse set of tightly orchestrated mechanisms to ensure clearance of mRNA from the cell.

Whilst the majority of mRNA degradation happens in the cytosol, decay also occurs in the nucleus. Again, the paradigm of either 5'→3' or 3'→5' digestion by exoribonucleases applies. 5'→3' digestion in the nucleus is done by XRN2, a member of the same protein family as the cytosolic XRN1 (Nagarajan et al., 2013). XNR2 acts co-transcriptionally on long transcripts (greater than 1,000nt) that lack a protective 5'-cap, typically due to the previous activity of an endonuclease (Rambout and Maquat, 2024). 3'→5' digestion, again similar to the cytosol, is done by the nuclear RNA exosome complex. There are a range of adaptor proteins that target pre-mRNA and mRNA with specific features to the exosome complex, with two of the most prominent being NEXT (which targets transcripts with no poly(A)-tail) and PAXT (which targets transcripts with a poly(A)-tail) (Figure 1.1) (Rambout and Maquat, 2024).

The relative importance of these different degradation pathways varies between mRNA species, cell types and stress conditions. Some mRNAs are primarily decapped after deadenylation, whilst others are more prone to exosome-mediated decay. mRNA 3'UTRs often contain specific sequence elements that regulate stability by recruiting RNA-binding proteins (RBPs) or micro RNAs (miRNA). For example, many inflammatory cytokine mRNAs have AU-rich elements (AREs) in their 3'UTRs that recruit RBPs which in turn recruit deadenylases and decapping enzymes, thereby accelerating decay of those mRNAs and limiting cytokine production (Chen and Shyu, 1995; Kratochvill et al., 2015). Conversely, certain RBPs bind 3'UTRs and enhance stability (e.g. HuR can stabilise some growth-related mRNAs under stress) (Brennan and Steitz, 2001).

1.3 Modelling mRNA turnover

Whilst both transcription and degradation are complex processes, they can be effectively represented in a simple kinetic model (Jürges et al., 2018). A common framework assumes that mRNA is synthesised at a constant rate and decays via a first-order process. Let $M(t)$ denote the concentration (or number) of mRNA molecules at time t. The rate of change of $M(t)$ can be described by the differential equation

$$\frac{dM(t)}{dt} = \sigma - \delta M(t), \tag{1.1}$$

where σ is the constant rate of mRNA synthesis (transcription) and δ is the decay rate constant, representing a first-order degradation process in which an mRNA has a certain probability of being degraded at any given time point. At steady state (SS), where $\frac{dM(t)}{dt} = 0$, the mRNA concentration is

$$M_{\text{ss}} = \frac{\sigma}{\delta}. \tag{1.2}$$

The half-life, $t_{1/2}$, of the mRNA is defined as the time required for the mRNA concentration to decrease to half of its initial value when synthesis is halted (i.e. when $\sigma = 0$). In that case, Equation 1.1 simplifies to

$$\frac{dM(t)}{dt} = -\delta M(t), \tag{1.3}$$

with the solution

$$M(t) = M_0 e^{-\delta t}. \tag{1.4}$$

Setting $M(t_{1/2}) = \frac{M_0}{2}$ (the definition of the half-life) leads to

$$\frac{1}{2} = e^{-\delta t_{1/2}} \quad \Longrightarrow \quad t_{1/2} = \frac{\ln 2}{\delta}. \tag{1.5}$$

This model provides a simple yet powerful framework to understand mRNA dynamics in steady state and under perturbations such as transcriptional inhibition (Rummel et al., 2023). In steady-state conditions where transcription and decay are balanced, the age distribution of the mRNAs exponentially decreases, meaning many transcripts are 'young' and relatively few survive to old age. Given exponential decay, the half-life is also directly related to the expected age (or 'mean age') of the mRNA by

$$\bar{\tau} = \frac{t_{1/2}}{\ln 2} \tag{1.6}$$

where $\bar{\tau}$ is the mean age of the population of mRNA.

1.3.1 Measuring mRNA kinetics

The main mechanisms that a cell has at its disposal to change the amount of mRNA available for protein translation are to alter transcription, σ, or degradation, δ, rates. As such, measuring σ and δ is of interest, both at steady state and in response to stimuli. Measuring the half-lives of human mRNA typically uses either transcriptional arrest (Figure 1.2A) or metabolic labelling (pulse-chase or approach to equilibrium (Figure 1.2B)) (Lugowski et al., 2018). In transcriptional arrest methods, a chemical such as Actinomycin D (which intercalates with DNA and prevents unwinding by DNA helicase) or α-Amanitin (which inhibits RNA Polymerase II directly) is used to halt the production of new transcripts(Chen et al., 1996; Bushnell et al., 2002). Samples are taken at increasing time intervals after transcriptional arrest and the degradation rate determined from the disappearance of the RNA over the time course. These experiments require careful normalisation of the sequencing data, typically with spike-in controls, and also suffer from some cellular stress

response to the chemicals and transcription halting (Lugowski et al., 2018; Viegas et al., 2023). As such, metabolic labelling offers some advantages.

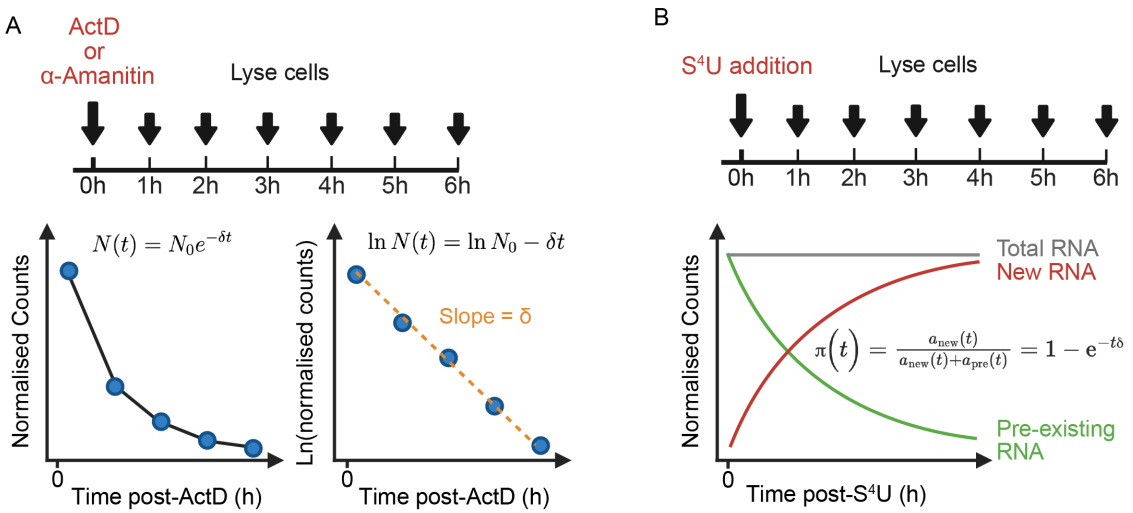

Figure 1.2: **Methods for estimating mRNA degradation and synthesis kinetics.** (A) Transcriptional arrest method use a chemical such as Actinomycin D (ActD) or α-Amanitin to halt transcription. This is followed by sequencing of samples at successive time points, normalisation of the abundances between time points to account for library size and RNA composition effects and finally model fitting to obtain an estimate of the half-life - assuming an exponential decay model of RNA degradation. The models used to estimate the parameters from abundance and log-transformed abundance data are shown at the top of the two plots. (B) Metabolic labelling methods with nucleotide analogous come in a variety of protocols depending on the experimental aims and the kinetics of interest. The change in the proportions of new RNA (i.e. labelled - shown in red) and pre-existing RNA (i.e. unlabelled - shown in green) following the change to 4sU media at $t = 0$ is shown. The equation shows how the proportion, π, of new RNA (a_{new}) to pre-existing RNA (a_{pre}) at any given time t directly relates to the degradation rate δ (Jürges et al., 2018). Other more advanced methods to estimate synthesis and degradation rates - including from single time point measurements - can be applied to metabolic labelling RNA-seq data using the grandR package (Rummel et al., 2023). (Created with biorender.com)

Metabolic labelling methods introduce a labelled precursor - such as 4-thiouridine (4sU) - that gets incorporated into newly synthesised RNAs, allowing new transcripts to be distinguished from old ones (Cleary et al., 2005; Dölken et al., 2008). In pulse-chase experiments, 4sU is added briefly (pulse) and then the media changed to normal uridine with samples taken at successive time points to assess the loss of 4sU over time (chase). In 'approach to equilibrium' set-ups, cells are cultured in analogue-containing media and samples taken at time points to measure the accu-

mulation of 4sU over time. Isolated RNA is treated to allow quantification of the new and old RNA transcript pools through protocols such as BRIC-seq (Imamachi et al., 2014), SLAM-seq (Herzog et al., 2017; Jürges et al., 2018; Rummel et al., 2023) and Timelapse-Seq (Schofield et al., 2018), with the latter two using chemistries that enable 4sU incorporation to be detected as T-to-C conversion events in RNA-seq data. Direct nucleotide conversion overcomes a laborious and error-introducing step in early metabolic labelling methods in which the labelled and unlabelled RNA fractions had to be separated Herzog et al. (2017). The proportion of new transcripts to total (or old) transcripts can be used to infer both the synthesis rate, σ, and the degradation rate δ (Figure 1.2B). Rummel et al. (2023) compare several statistical frameworks that can estimate these parameters from pulse-chase, approach to equilibrium and from single measurements (provided that there is a reference measurement before labelling). In particular, they demonstrate not only being able to estimate the steady state levels of synthesis and decay but also the log(fold-change) of the two parameters between experimental conditions. The key concept underpinning these methods is that the kinetics can be inferred without transcriptional arrest if new (i.e. young) RNA can be distinguished from old RNA - this is of particular relevance for Chapter 3, Section 3.5.

The typical mRNA half-life in human cells is thought to be in the range of 4-10 hours, such as in Rummel et al. (2023) where they found a median age of around 4 hours using 4sU labelling and 8 hours using Act-D transcriptional shut-off in a human lung epithelial cell line, highlighting the discrepancy arising from different measurement methods (Finkel et al., 2021). There is also a large degree of relevant biological variation in transcript ages: certain transcripts (often those encoding regulatory proteins like transcription factors or cytokines) are extremely short-lived with half-lives of under 1 hour, whereas others (housekeeping genes, structural proteins) can have half-lives exceeding 20 hours (Yang et al., 2003). Recently, machine learning models that predict half-life from mRNA sequence that are trained on ensembles of experimentally derived half-lives has improved rapidly (Agarwal and Kelley, 2022;

Yaish and Orenstein, 2022). However, questions remain as to the source of variation between measured half-lives from different studies, which machine learning models typically consider noise.

Whilst many degradation rates are estimated at steady state, modulation of degradation rates is an important process that enables cells to respond to various stimuli. Several studies that have separated the contributions of changes to transcription and degradation rates to stimulus indicate that changes to transcription drive the majority of the changes (Barenco et al., 2009; Rabani et al., 2011). There is also a growing body of evidence for the existence of feedback mechanisms between transcription and degradation rates (Braun and Young, 2014; Hartenian and Glaunsinger, 2019). A recent study applying RATEseq in Rett syndrome, a rare neurodevelopmental condition in which transcription rates are widely dysregulation, showed instances where mRNA half-life changes completely counterbalanced transcription changes to preserve the level of total mRNA of specific genes Rodrigues et al. (2023). Thus, the relative contributions of transcription and degradation rates deserve greater attention in interpreting levels of gene expression in RNA-seq studies, making the development of new methods that provide easy access to these underlying kinetics an area of substantial interest.

1.4 A-to-I RNA Editing by ADAR proteins

On its path from birth to death, an mRNA can be adorned with a large number of post-transcriptional modifications which together fine-tune its splicing, export, localisation, translation efficiency and decay (Roundtree et al., 2017; Xiang et al., 2018; Boo and Kim, 2020; Delaunay et al., 2023). Adenosine-to-inosine (A-to-I) RNA editing is one of the most common post transcriptional modifications in animals and is primarily catalysed by the adenosine deaminase acting on RNA (ADAR) family of proteins. Inosine is interpreted as guanosine by the cell's translation machinery and in sequencing library preparations and thus an I is read as G in both

contexts (Pestal et al., 2015). Three ADAR genes exist in humans: ADAR1, ADAR2 and ADAR3 - which share a common N-terminal dsRNA-binding domain and a C-terminal deaminase domain (Figure 1.3A) (Herbert et al., 1997; Stefl et al., 2006; Nishikura, 2015). ADAR3 is expressed in the brain and, whilst catalytically inactive, can still bind dsRNA and is thus thought to act as a negative regulator of editing by the other ADAR proteins (Mladenova et al., 2018; Raghava Kurup, Oakes, Manning, Mukherjee, Vadlamani and Hundley, 2022). Ongoing research continues to expand the functional roles for ADAR3, with recent studies implicating it in increasing NF-κB activation which can confer resistance to cancer drugs in glioblastoma models, as well as a role in murine neuronal development (Raghava Kurup, Oakes, Vadlamani, Nwosu, Danthi and Hundley, 2022; or Karlström et al., 2024).

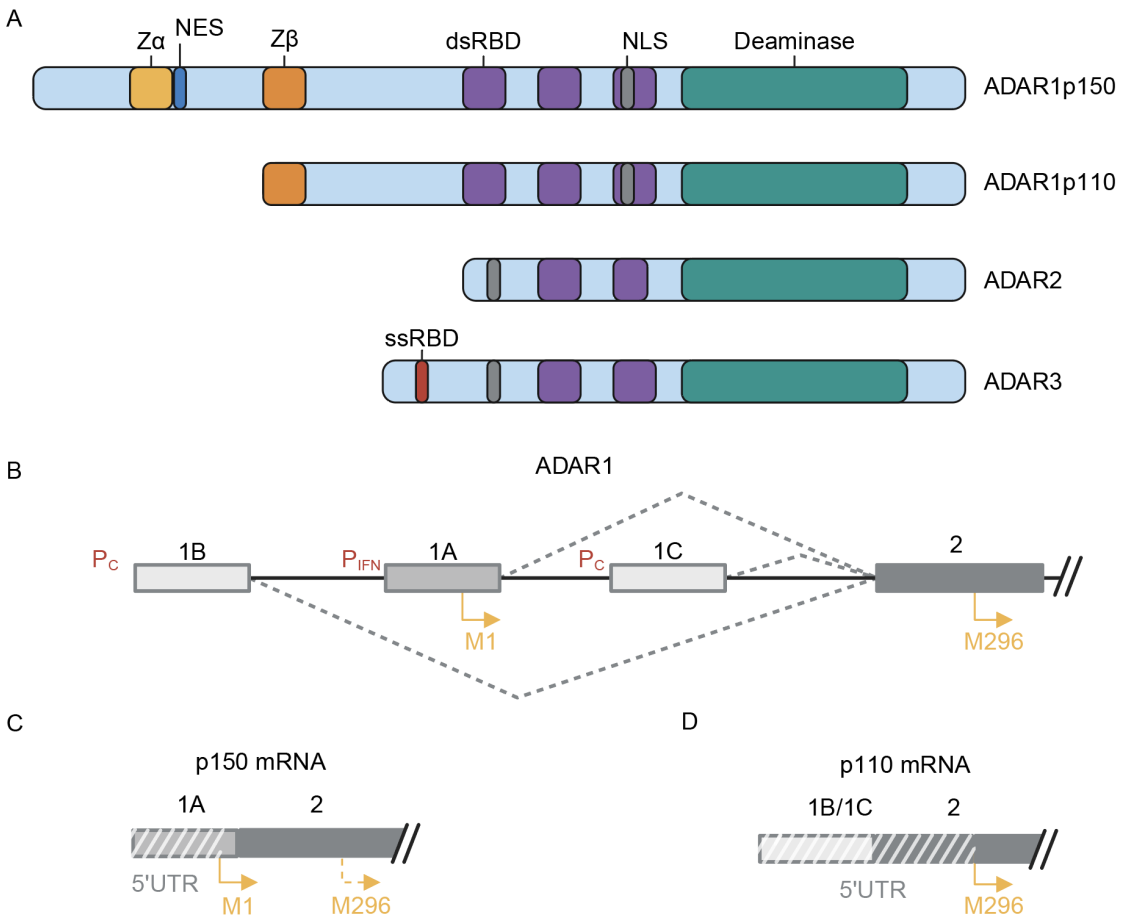

Figure 1.3: **Protein domain and gene structure of the human ADAR proteins.** (A) Schematic of the domains of the human ADAR1 (both p150 and p110 protein isoforms), ADAR2 and ADAR3. Labelled are the Z-RNA binding domains (Zα and Zβ), nuclear export signal (NES), double-stranded RNA binding domains (dsRBD), nuclear localisation signal (NLS), the deaminase domain and the single-stranded RNA binding domain (ssRBD) of ADAR3. The schematic is adapted from Nishikura (2015). (B) The potential splicing outcomes of the start of the ADAR1 gene. The boxes show exons: namely the three optional first exons (1B,1A,1C) and exon 2. The positions of the two constitutive promoters in front of exon 1B and exon 1C are shown, as well as the interferon-inducible promoter of exon 1A. The yellow arrows indicate methionine start codons (M1 and M296) that produce the p150 and p110 protein products, respectively. Dashed arrows indicate potential splicing processes. (C & D) show the outcomes of splicing that produce the start of the p150 mRNA (C) and the p110 mRNA (D). Joined exons are labelled and coloured as in (B). The dashed area in (C) shows the region of exon 1A that becomes part of the 5'UTR of the p150 transcript. The M296 start codon (dashed yellow arrow) has leaky translation which can produce the p110 protein. (D) Irrespective of whether exon 1B or 1C are joined to exon 2, both - along with a substantial amount of exon 2 - become part of the 5'UTR of p110. The Zα domain of p150 is encoded in exon 2, upstream of M296 and thus is absent from the p110 protein product. Schematics are adapted from Lykke-Andersen et al. (2007) and Sun et al. (2021) and created with biorender.com

ADAR2 is ubiquitously expressed but with elevated levels in the brain. It is most commonly associated with edits in coding regions (so called 'recoding events') that result in amino acid substitutions, such as a critical edit on the mRNA of GluA2 - part of the AMPA receptor. This specific edit introduces a Q/R codon change that affects ion channel permeability and the loss of which is implicated in disease (Higuchi et al., 1993; Salpietro et al., 2019).

ADAR1 is responsible for the vast majority of editing events in humans (thought to be on the order of millions of sites) and is expressed in all tissues (Bazak et al., 2014; Tan et al., 2017; Schaffer et al., 2020). In humans, ADAR1 is present as two isoforms: the full-length, (mostly) interferon-inducible p150 isoform and the (mostly) constitutively expressed p110 isoform, which lacks one of the Z-RNA binding domains of p150 (George and Samuel, 1999; Sun et al., 2021) (Figure 1.3A,B,C). Whilst p110 is mostly localised in the nucleus (and in particular, to the nucleolus) p150 contains an N-terminal nuclear export signal and is largely found in the cytosol, although it can shuttle between the two compartments (Poulsen et al., 2001; Desterro et al., 2005). Recent research has shown that the p150 mRNA is capable of producing both the p150 and p110 protein isoforms due to leaky ribosome scanning (Figure 1.3C), making both isoforms interferon-inducible (Sun et al., 2021).

ADAR1 is known to play a key role in preventing autoimmunity to host-derived dsRNA. The human genome contains many repeat sequences that form long dsRNA regions when expressed. The most common of these sequences are *Alu* elements, a type of short interspersed nuclear element (SINE) that are roughly 300nt in length and exist at over a million copies in the human genome, accounting for roughly 10% of the entire sequence (Lee et al., 2024). *Alu* elements are enriched in introns and 3'UTRs and can form dsRNA when two are found close together but in opposite orientations. Roughly 80% of all ADAR1 edits are thought to occur in *Alu* sequences (Sun et al., 2021).

Cytosolic double-stranded RNA is sensed by at least two distinct antiviral mech-

anisms. First, long dsRNA is recognised by the RIG-I-like receptors RIG-I and MDA5, which signal via the mitochondrial adapter MAVS to induce a type I interferon response (Rice et al., 2012; de Reuver et al., 2022). Second, dsRNA can activate the eIF2α kinase PKR (EIF2AK2), leading to translational arrest and further activation of type I interferon response. In humans, loss of ADAR1 editing causes the severe interferonopathy, Aicardi-Goutières syndrome, characterised by constitutive activation of these pathways and widespread inflammation (Rice et al., 2012; de Reuver et al., 2022).

Recent work has shown that these two axes are kept in check by the cytosolic ADAR1-p150 isoform (Figure 1.1). Its editing activity converts endogenous *Alu*-derived duplex RNAs into inosine-containing strands that fail to activate MDA5, thereby supressing MAVS-dependent interferon production (Liddicoat et al., 2015; Hu et al., 2023). Independently, the Zα domain of ADAR1-p150 binds Z-conformation RNA and blocks both ZBP1-mediated cell death (via apoptotic and necroptotic routes) and PKR activation by unedited duplexes (Karki et al., 2021; de Reuver et al., 2022). In particular, unedited *Alu* duplexes are strong ZBP1 agonists, triggering cell death when not bound by ADAR1 (de Reuver et al., 2022). Thus, whilst it was previously thought that most ADAR1 editing happens in the nucleus, recent findings stress the importance of cytosolic ADAR1-p150 editing in suppressing autoimmunity (Deng et al., 2025).

1.5 RNA timestamping

Rodriques et al. (2020) first combined the concepts of A-to-I editing and measuring gene expression through the framework of RNA timestamping. Since many of the concepts from their work underpin the work and concepts in this thesis, it is described in detail below.

The authors asked whether the ages of individual RNA molecules - that is, the time since a given molecule was transcribed - could be determined by the number of A-to-

I edits observed on individual sequencing reads. It was hypothesised that the longer an editable RNA molecule exists in the cell, the more time it has to accumulate A-to-I edits (akin to carbon dating of rock samples by measuring the outcomes of radioactive decay). To create such a system, they inserted a custom adenosine-rich RNA cassette (a 'timestamp') into the 3'UTR region of a gene of interest that, when expressed from a plasmid, would form dsRNA with stem-loop structures. They also engineered an ADAR2-MCP fusion protein which would be recruited to the stem-loop structures of the timestamp by its MCP domain and edit the surrounding area (Figure 1.4A). The accumulation of the A-to-I edits on individual transcripts (i.e. single RNA molecules) could be determined by RNA-seq where edits appeared as A-to-G mutations (Figure 1.4B). They determined the rate at which A-to-I edits accumulated at each position on the timestamp through a 'calibration' experiment, which involved expressing the timestamped gene by adding doxycycline, halting production of new transcripts with ActD, sequencing at increasing time intervals and measuring the fraction of edited reads mapping to each editing site. Since there was no new transcription (which would otherwise introduce new adenosines into the system) and since the degradation rate of edited and unedited transcripts was assumed to be the same, the fraction of adenosines at each editing site was determined solely by the rate at which the ADAR2 construct edited that position. In the sequencing data, this was observed as an increase in the editing level at each time point taken post-ActD (Figure 1.4B).

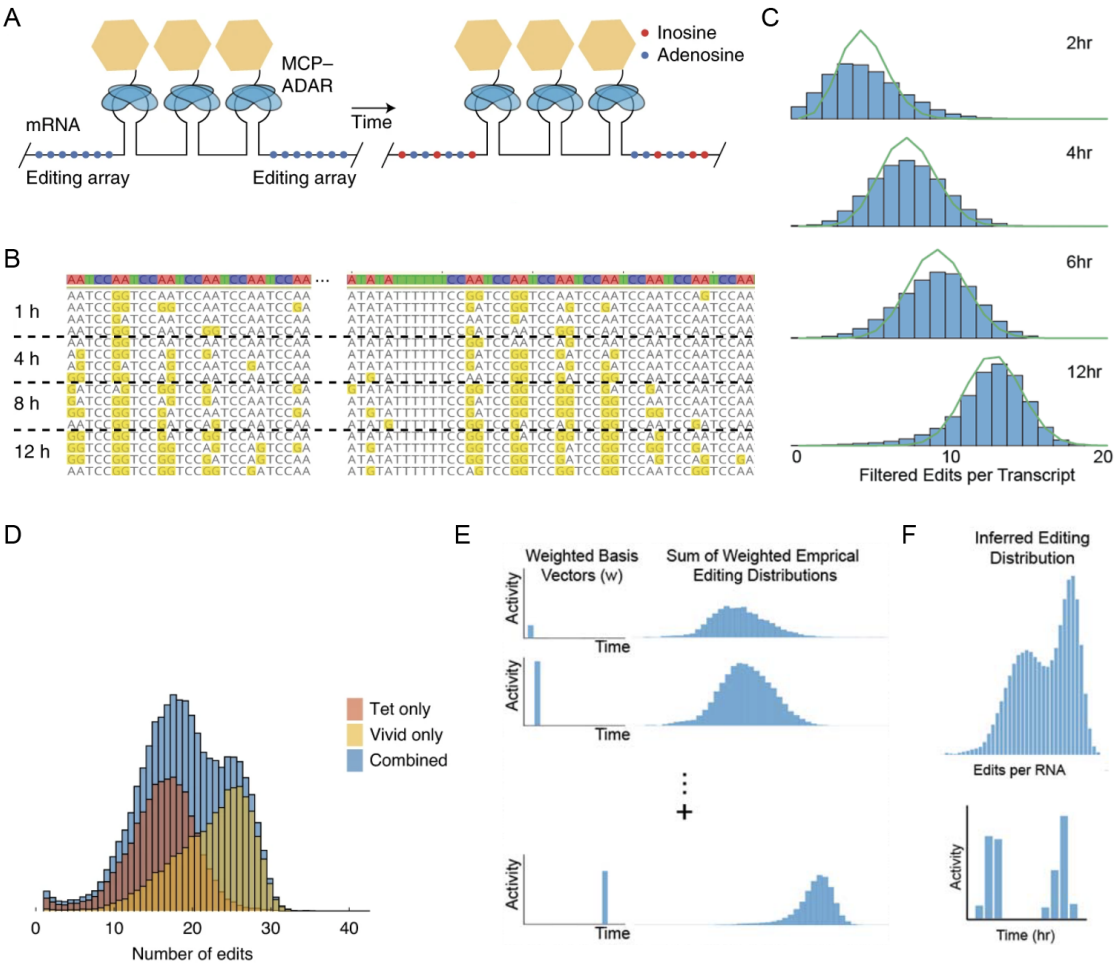

Figure 1.4: **A-to-I edits encode time on RNA timestamps.** All panes are taken from Rodriques et al. (2020) with permission of the author. (A) Schematic depicting the recruitment of ADAR2 to an RNA timestamp by a fused MCP domain. (B) The increase of edits (guanosines, highlighted in yellow) on multiple copies of a timestamp in RNA-seq data 1h, 4h, 8h and 12h after transcriptional halting. (C) Blue shows the histograms of the number of edits observed per transcript over multiple copies. The green line shows the Poisson binomial distribution when the value of t is the time since doxycycline was added. (D) Histograms of edits per transcript for barcoded RNAs carrying the same timestamp cassette but expressed from two different promoters (Tet and Vivid) 4 hours apart. The red histogram shows the transcripts generated from the Tet promoter and the yellow histogram shows those from the Vivid promoter. The blue histogram shows the combined histogram, from which two peaks can clearly be seen. (E & F) The gradient descent method for determining the most likely transcriptional programme that produced the observed editing histogram. (E) Shows the 1 hour basis vectors (left) and the corresponding editing histograms (right) that would be generated from that vector. (F) Gradient descent is used to determine the weighted combination of basis vectors that best describe the observed editing distribution of edits per RNA. Panes A and B are reproduced from Fig. 1 of Rodriques et al. (2020), pane C is from Supp. Fig. 3, pane D is from Fig. 3, panes E and F are from Supp. Fig. 8.

Mathematically, the authors modelled the editing by the ADAR2 construct at a site, i, as a Bernoulli trial where a 'success' event (i.e. observing an edit, G, in the sequencing data) has a fixed probability per unit time. Since many reads map over any given site, the observed fraction of edited reads y at site i at time t is given by

$$y_i(t) = 1 - e^{-\lambda_i t}. \tag{1.7}$$

In the calibration experiment, since t is known and $y_i(t)$ is observed, λ_i can be determined by fitting the equation to the accumulation of edits over the duration of the experiment[1]. Using this framework, the authors determined the editing rates, λ_i, for all editing sites and kept those that were well fit (coefficient of determination, R^2, was greater than 0.9). They found that the number of observed edits per read was highly correlated between biological replicates, indicating that the editing rates were robust.

Having determined the editing rates, they devised a model to infer when gene expression was induced from the observed number of edits. For this they calculate the probability of observing exactly n edits at time t using a Poisson binomial model of the form

$$p\left(n,t\right) = \sum_{\mathrm{A:sum}(A)=n} \prod_{k:\mathrm{A_k}=1} y_k(t) \prod_{j:\mathrm{A_j}=0} 1 - y_j\left(t\right), \tag{1.8}$$

where A is a binary vector that gives the observed edit state (1 if edited, 0 if unedited) of each adenosine on the timestamp. By comparing the observed number of edits, $q(n)$, to the model, $p(n,t)$ they estimate the time, t, that the transcript was produced (Figure 1.4C). In practice, since multiple copies of the same transcript are produced during transcription, they compare the distribution of the number of edits over all the copies of a given transcript with $p(n,t)$. As expected, having more copies of a given transcript decreases the error of the estimated time since the promoter

[1]In practice, the authors fit a modified version of equation 1.7 to the RNA-seq data which accounted for the gap in time between Dox addition and ActD addition

was activated.

The authors demonstrate that using this framework, they can determine when the promoter driving the gene of interest was activated. They do this by generating the $p(n, t)$ distributions for each hour between $t = 0$ hours and $t = 12$ hours to give a set of 12 'basis vectors' (Figure 1.4E). They then use gradient descent to optimise which combination of these basis vectors best fits the observed distribution of edits per transcript (Figure 1.4F). They show that this can discern between two transcriptional pulses (using the same timestamp but driven by two different promoters at different times (Figure 1.4D)) and can order single cells by when the promoter was activated.

The authors' framework, however, has several features that would make it challenging to adapt beyond their highly engineered system. First, each gene of interest needs to be genetically engineered to insert a timestamp in its 3'UTR, which makes it unscalable. Second, their framework only answers questions of when a given promoter was turned on and does not answer questions about synthesis and degradation rates of the transcripts. Third, the method of comparing the distribution of the number of edits per transcript with the Poisson binomial model is at the same time cumbersome and over-simplifying since it only considers the number of edited sites, n, rather than *which* sites are edited. Improvements to these shortcomings are introduced in Chapter 2 and discussed at length in Chapter 3.

1.6 Advances in RNA-sequencing

1.6.1 Short-read sequencing

The underpinning technology of the previous sections is RNA-sequencing, which has enabled a transformation in our ability to quantify the RNA present in cells. The late 1900s and early 2000s saw competition between several sequencing technologies, including Sanger sequencing (Sanger and Coulson, 1975), microarrays (Schena et al.,

1995; Clark et al., 2007), and sequencing by synthesis (Nagalakshmi et al., 2008; Lister et al., 2008; Bentley et al., 2008; Wang et al., 2009). Whilst microarrays were the dominant platform for measuring gene expression in the early and mid-2000s, Illumina's sequencing-by-synthesis technologies rose to prominence in the 2010s following the acquisition of Solexa in 2007 and the release of the HiSeq sequencing machine in 2010 (Barba et al., 2014).

Illumina sequencing requires DNA as input and thus the mRNA from a sample must be converted into complementary DNA (cDNA) before flow cell loading. There are many different library preparation techniques to obtain cDNA from input RNA, the most common of which is TruSeq (Ura et al., 2022). In this protocol, mRNA is isolated by using oligo-dT magnetic beads that bind to the poly(A)-tails of the mRNA through Watson-Crick base pairing and enable separation. The mRNA are then fragmented into approximately 200nt fragments and a first complementary DNA strand synthesised by reverse transcriptase (RT). To generate a stranded library, the second cDNA strand is synthesised using dUTP instead of dTTP, which can be digested by the USER enzyme. Following adaptor ligation, the cDNA library can then be amplified by polymerase chain reaction (PCR) to obtain the required concentration for loading onto the flow cell of choice.

Sequencing on the flow cell begins with bridge amplification of the cDNA library, which creates clonal clusters on the flow cell through successive amplification steps (PCR reactions). Sequencing by synthesis then adds bases that are covalently attached to a fluorophore, which emits a characteristic colour upon incorporation which is captured by a high performance camera. The light emitted at millions of clusters is captured in parallel at every sequential cycle (typically for 150 cycles) and the total run is output as a .bcl format file, which contains information on the base call at each position in the read and an accuracy metric - a Q-score - which is Phred scaled ($Q = -10\log_{10}(P)$), where P is the error probability. The bcl files are converted to FASTQ format, which can then be aligned against a reference genome or transcriptome (if desired) resulting in .bam file that contains details of

the alignment.

Whilst short-read sequencing is cheap and widely available, the length of the reads is such that only a portion of a given transcript is captured on a single-read. This presents two major issues: the first is resolving isoforms, since unless a read spans an exon-intron junction then it is impossible to know which other exons and introns were present on the original transcript that the read is mapping to. The second, related, issue is 'phasing' of single nucleotide variants on the transcripts, such as for determining how much A-to-I editing has taken place on individual transcripts. For endogenous ADAR1 editing, the separation between editing sites is large enough that only a few editing sites can be captured (or 'phased') on a single short-read. In RNA timestamps as detailed by Rodriques et al. (2020) it is essential to phase all of the edits from each timestamp, which is possible since the timestamps themselves are short RNA sequences. However, phasing all of the possible editing sites on endogenous transcripts generally requires long-read sequencing.

1.6.2 Long-read sequencing

The two major technologies for long-read sequencing are developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Both companies were founded prior to Illumina's acquisition of Solexa (2004 for PacBio, 2005 for ONT). However, achieving comparable accuracy profiles and ease of use at a competitive price point with Illumina has proved challenging.

PacBio offers long-reads on the order of 10-20kb and median per-read accuracies of Q30 or more through circular consensus sequencing (CCS) (Wenger et al., 2019). In CCS, cDNA is circularised and immobilised in a single micro-well, around which a polymerase processes and the sequence read out optically using fluorophores. As the polymerase makes multiple passes, a consensus of the sequence of the circular cDNA can be constructed, producing high fidelity (Hi-Fi) reads. PacBio has traditionally been expensive and throughput limited (a Sequel IIe sequencer with a SMRT Cell 8M

yields roughly 2-5 million reads), although the advent of the MAS-ISO-seq protocol and Revio sequencer has recently improved both of these metrics (Al Khafaji et al., 2023).

ONT offers exceptionally long-reads that can exceed a million base pairs and high throughput from simple library preparations. Nanopore sequencing involves threading a single cDNA strand through a highly engineered, membrane embedded protein nanopore. As the DNA strand passes through the pore, it changes the electrical conductivity of the pore, which can be measured by an electrode. The resulting trace of the voltage change over time at each pore is decoded by machine learning algorithms called 'basecallers' that infer the sequence of the molecule and an associated Q-score (MacKenzie and Argyropoulos, 2023). Traditionally, the accuracy of ONT sequencing has been poor but with the introduction of the R10.4.1 chemistry in 2022, median per-read Q-scores are in excess of Q20. Most ONT sequencing is done on either a MinION flow cell (which typically outputs ∼10-15 million reads) or a PromethION flow cell (which typically outputs 60-90 million reads) and both can be highly parallelised on the gridION and PromethION24/48 machines, respectively. It is also possible to do direct RNA-sequencing rather than cDNA with ONT, however this uses a different nanopore which has a lower translocation rate of the RNA strand and so the yield substantially lags that of cDNA sequencing (Nguyen et al., 2022).

Both ONT and PacBio offer reads that are longer than the typical length of a human mRNA transcript (mean = 3.5kb, SD = 2.5kb (Piovesan et al., 2019)) and can capture the sequence of entire transcripts and overcome the isoform and phasing barriers described for short-read sequencing. Whilst it is generally assumed that ONT has much higher error rates than Illumina or PacBio, a recent study suggests that the error introduced by polymerases in both Illumina (bridge amplification) and PacBio (circular consensus sequencing) methods are underappreciated (Sun et al., 2024). The Q-scores reported in the FASTQ files denote the error of correctly identifying the nucleotide being *sequenced*, not the nucleotide that was *present* on

the input RNA. Polymerase errors can still be assigned a high Q-score even though the sequence itself contains an error with respect to the starting RNA. Therefore, the true error rates of Illumina and PacBio are higher than reported, although the size of the discrepancy depends not only on the number of PCR cycles in the sequencing itself but also in the library prep.

1.6.3 Single-cell sequencing

Over the past decade, the ability to profile transcriptomes at single-cell resolution has transformed our understanding of cellular heterogeneity. The first genome-wide single-cell RNA-sequencing (scRNA-seq) experiments appeared in 2009, when Tang and colleagues demonstrated that full-length mRNA could be reverse-transcribed and amplified from individual mouse blastomeres (Tang et al., 2009). Soon after, plate-based protocols - such as Smart-Seq (Ramsköld et al., 2012; Picelli et al., 2014; Hagemann-Jensen et al., 2020) - improved sensitivity and broadened coverage - enabling the detection of thousands of genes per cell.

In May 2015, two seminal papers were published in the same edition of *Cell* that used microfluidics and droplet-encapsulation to massively increase the number of cells that could be sequenced with scRNA-seq. Drop-seq (Macosko et al., 2015) and InDrop (Klein et al., 2015) both encapsulate individual cells together with barcoded beads (or hydrogel spheres) in nanolitre-scale droplets, allowing tens of thousands of cells to be profiled in a single experiment. These technologies were rapidly commercialised, such as in the 10x Genomics Chromium system which has led to a surge in the use of scRNA-seq in biomedical research.

The most widely used methods now fall into two broad classes: plate-based and droplet-based. Plate-based approaches (e.g. Smart-Seq2 (Picelli et al., 2013)) still offer greater per-cell sequencing depth and full-length coverage, making them well suited to studies of isoform diversity or lowly expressed genes. Droplet-based techniques (e.g. Drop-seq (Macosko et al., 2015)) excel at sampling millions of cells from

complex tissues, enabling the discovery of new cell types and trajectories at scale.

The most popular method, the 10x Genomics Chromium 3' assay, primes exclusively at the poly(A) tail (after oligo-dT capture) and only captures a small portion of the 3' end of the sequence. In practice, the first read encodes the cell barcode and a unique molecular identifier (UMI), while the second read sequences just the 3'UTR and/or the terminal exon. This is enough to identify which gene the transcript originates from but information on the rest of the sequence (such as alternative splicing, single nucleotide variants, etc.) are lost. Two recent developments improve read coverage over the full length of the input RNA. Smart-seq3 (Hagemann-Jensen et al., 2020) uses a template switching oligo, that contains a UMI, at the 5' end of the transcript. By doing a small amount of PCR, tagmentation with Tn5 and paired-end short-read sequencing, many paired end reads will share the same UMI-containing first read and a second read that maps somewhere further from the 5' end into the sequence. Computationally, all of the second reads can then be linked since they share the same UMI on the first read of the pair. Alternatively, VASA-seq (Salmen et al., 2022) performs fragmentation within droplets, enabling the cell-identifying adaptors to be ligated to every fragment, which has shown improved uniformity of coverage and sensitivity over Smart-seq3. Still, however, partial coverage of the original RNA and the limitations of short-read sequencing detailed above (e.g. isoforms, phasing), persist.

Long-read single cell sequencing is still in its infancy and suffers from low yield (De Jonghe et al., 2024). It is, however, an attractive prospect since using either PacBio or (in particular) ONT in combination with high yielding droplet-based single cell methods would enable the sequencing of full transcripts in thousands or potentially millions of single cells. At present however, plate-based long-read single cell sequencing tends to provide higher yields and so is preferred for isoform analysis, even if fewer cells are sequenced. The error rate of ONT sequencing, even with the Q20+ R10.4.1 chemistry, presents a particular challenge for counting transcripts with UMIs, which are typically 6-12 bases long and which must be sequenced with

perfect accuracy to function as intended. Several methods have been developed that error-correct UMIs with new chemistry, *in silico* or both (Lebrigand et al., 2020; Philpott et al., 2021). Whilst high-yield, single-cell long-read sequencing is difficult to perform, it is possible and allows for phasing of all A-to-I edits on individual transcripts.

1.7 Towards *in vivo* measurement of gene expression dynamics

Whilst the methods for measuring the dynamics of gene expression have advanced substantially in the last decade, no method exists that combines the endogenous property of RNA velocity analysis, with the true-time and high-resolution kinetic measurements of metabolic labelling. In order to move towards *in vivo* measurements in humans, a new label free method is required. In this thesis I present a new method - Endogenous RNA Age (ERA) - which enables label-free measurement of gene expression dynamics in human cells. Conceptually, the method implements the principle of metabolic labelling - that nucleotide changes can discern old transcripts from young transcripts - but using the endogenous process of A-to-I editing as the label. Whereas metabolic labelling is only able to discern between transcripts that have been produced since the label was added and those that were already present, ERA estimates the time at which individual transcripts were transcribed. Building on recent advances in long-read and single-cell sequencing, the method can estimate the mean ages of hundreds of genes with prevalent A-to-I editing from RNA-sequencing data and detect past changes in transcription - which is showcased in primary human monocytes.

1.8 Aims & Thesis Structure

The motivation for this research arose directly out of the RNA timestamps paper of Rodriques et al. (2019). The co-first author of that paper, Sam Rodriques, was one of the supervisors on this project. The research in this thesis sought to answer two overarching questions:

1. Are the rates of endogenous A-to-I editing events sufficient to enable timestamping of endogenous transcripts?

2. If so, what can the ages of endogenous transcripts tell us about gene expression changes?

I answer these questions over three core chapters, which are summarised below:

1. **The First Core Chapter (Chapter 2)**. This Chapter focuses on endogenous A-to-I editing and introduces the theoretical framework of editing rates. It is already known ADAR1 and ADAR2 are expressed ubiquitously across human cell types and that they edit at millions of loci. However, it is not known what the underlying editing rates are at these loci nor is there a methodology to determine them. This Chapter reports an adaptation of the calibration protocol introduced in RNA timestamps (Rodriques et al., 2020), which is used to produce the first, transcriptome-scale determination of A-to-I editing rates in human cells. Several calibration datasets were produced in a variety of human cell types and the correlation of editing rates between them determined. The editing sites are characterised by their editing rates, where they are located in gene bodies (e.g. intron, 3'UTR, exon, etc.) and how they are distributed over the expressed genes in cells. This analysis is continued in the second core chapter after the per-site age method is introduced.

2. **The Second Core Chapter (Chapter 3)**. This Chapter tackles the question of whether endogenous A-to-I edits can be used to infer the ages of endogenous transcripts. The framework of Rodriques et al. (2020) could not be

directly used in the endogenous setting for several important reasons. First, single-molecule age estimation requires phasing of edits. In RNA Timestamps, this was possible with short-read sequencing since the timestamp RNA cassettes were short. However, editing sites on endogenous transcripts occurs over far greater lengths of sequence and so long-read sequencing was required. Second, the mathematical model of RNA timestamps (namely the Poisson binomial, Equation 1.8) did not translate to the endogenous setting where genes are constantly being expressed. To solve these issues, I develop two methods for determining Endogenous RNA Age (ERA): per-site ERA, and per-transcript ERA. Per-site ERA estimates the mean age of a pool of transcripts, such as all the transcripts arising from a particular gene. It has the advantage of working for both short-read and long-read sequencing data. Per-transcript ERA estimates the age of individual transcripts from long-read sequencing data. I developed a novel likelihood-based framework for estimating the ages of individual reads based on their editing profiles. I compare the performance of the two methods on various short and long-read datasets and pre-processing pipelines. I conclude the chapter by showing that both per-site and per-transcript ERA can provide estimates of mRNA half-lives. Together, the first two chapters answer the first major research question.

3. **The Third Core Chapter (Chapter 5)**. This Chapter explores the second research question: 'what do the ages of transcripts tell us about biology?'. The first aim was to identify increases and decreases in gene expression in a highly-controlled system of three genes placed under inducible control (Tet-On). Rodriques et al. (2020) showed that the timing of promoter activation in the Tet-On system could be estimated from transcript ages. However, it was not known how this would work with the new models described in Chapter 3 nor whether decreases in promoter activity could also be measured. The second aim was to measure gene expression changes from endogenous regulation, which was investigated in HEK293 cells responding to heat shock. The

third aim was to identify gene expression changes in primary human cells, which is done by measuring the response of monocytes to lipopolysaccharide (LPS) treatment. The final aim was to apply ERA to single cell data, which is achieved for plate-based long-read sequencing of 96 HEK cells.

The thesis concludes with a discussion of the research undertaken and the results obtained. I put into context the newly characterised kinetics of endogenous A-to-I editing, the development of label-free computational methods to measure gene expression and how these contribute to the RNA-seq toolbox for understanding human biology.

1.8.1 Definition of terms

Throughout this thesis I use several terms as short-hand which I define clearly below:

- *Transcript* or *individual RNA molecule*: I use the terms 'transcript', 'individual RNA molecule' and 'individual mRNA' to refer to a single molecule of mRNA. For instance, when estimating the ages of individual transcripts, I am referring to a single mRNA transcript.
- *Mean gene age*: any and all references to 'gene age' refer to the age of the pool of transcripts that map to the same gene ID. I am not referring to the DNA sequence of the gene. All of the sequencing data here were aligned against the reference genome rather than transcriptome so different isoforms of the same gene are not considered.
- *Read*: When talking about sequencing reads and transcripts, I often use the two interchangeably - especially in the case of long-read sequencing. When I talk about the age or editing profile of a transcript, I am making the assumption that the long-read that I am analysing has a 1:1 correspondence to the transcript whose sequence it represents.
- *MLE, transcript age*: All transcript ages are inferred by maximum likelihood estimation (MLE) and so the terms 'MLE' and 'transcript age' are used inter-

changeably.

- *My collaborators*: This research project was a highly collaborative undertaking between the Mathematical Immunology Group at the University of Oxford (Professor Mark Coles and Professor Eamonn Gaffney) and the Applied Biotechnology Lab at the Francis Crick Institute (Dr. Sam Rodriques). All of the experimental work of this thesis (apart from the monocyte isolation, which was performed by Gabrielle Chappell) was undertaken in the lab at the Crick by either Dr. Ali Ghareeb or Dr. Aaron Wagen, two of my collaborators in the Applied Biotechnology Lab. Unless my collaborators (be that Ali, Aaron or other specific individuals) are mentioned explicitly, all of the work undertaken is my own. In sections where we were working very closely together, I write in the active voice to provide further clarity.

Chapter 2

Measuring Endogenous Editing Rates

"Age before beauty"

– Anonymous

2.1 Introduction

As introduced in the Chapter 1, Adenosine-to-Inosine (A-to-I) editing by ADAR1 is one of the most common mRNA modifications in humans and is found in all tissues. These edits can be detected as A/G mismatches in RNA sequencing data and can occur either co-transcriptionally, in the nucleus or in the cytosol. ADAR1 is expressed as two protein isoforms: p150 and p110. p150 contains an additional Z-RNA binding domain (Zα) compared with p110, is mostly found in the cytosol (whereas p110 is mostly found in the nucleus) and is thought to be the dominant isoform (Sun et al., 2021). Both isoforms display low sequence specificity, enabling them to edit potentially up to 100 million editing sites in humans, although the average reported number for any given cell type tends to be in the low millions (Bazak et al., 2014). ADAR1 edits in double-stranded RNA regions and shows a preference for inverted *Alu*-repeats, most commonly found in 3' untranslated regions (3'UTRs) and introns (Sun et al., 2021).

A-to-I editing has a broad range of biological consequences in humans and other animals. The majority of editing by ADAR1 is thought to prevent aberrant immune activation by cytosolic dsRNA, which the innate immune system can mis-identify as being viral in origin - precipitating a type I interferon response (Rice et al., 2012;

Liddicoat et al., 2015; Karki et al., 2021; de Reuver et al., 2022; Hu et al., 2023; Deng et al., 2025). Thus, the continuous activity of ADAR1 marks host dsRNA regions as self and maintains homeostasis.

RNA-seq analysis of many different disease states - in particular cancer and neurodegenerative/neurodevelopmental - has discovered altered editing profiles in these conditions, with many of the studies linking editing dysfunction to disease phenotype (Chan et al., 2020). As such, there is substantial interest from a biomedical perspective in accurately measuring and modelling editing sites from RNA-seq data.

Many methods have been developed to identify editing sites and quantify the level of editing at those loci. The key innovations are largely around controlling type II errors that may arise from mislabelling sequencing errors or genomic variants as edits (Piechotta et al., 2017, 2021). The two main tasks for quantifying editing are to: 1. detect editing sites in a single sample or 2. determine whether a site is differentially edited between two conditions. The information used in detecting an editing site is the number of edited reads that align to a putative site, and the total number of reads at that position: the ratio of the two gives the 'editing level'. In addition, models can now incorporate sequence features that may affect the accuracy of the RNA-seq and subsequent alignment including proximity to a splice site, being within a homopolymeric sequence or at the ends of reads Piechotta et al. (2017). However, the two basic thresholds imposed are a minimum number of reads at a given site (for instance, five reads), and a minimum editing level (for instance, one read edited read).

Whilst the measured editing level at a particular site is simply the ratio of edited reads to total reads, it is in fact the product of the mRNA dynamics (synthesis, processing, decay) and the enzymatic activity of ADAR at that site (i.e. the 'editing rate'). As a nascent transcript emerges from RNA polymerase II during transcription, all of its editing sites are unedited - it is only as the transcript 'ages' (i.e. has existed for longer in the cell) that its adenosines begin to be deaminated. However,

the distinction between the observed *editing level* and the underlying *editing rate* has been largely neglected, and the rates themselves have not been characterised at the transcriptome scale.

This Chapter explores endogenous editing rates through the following research questions:

1. **Do endogenous editing levels increase when transcription is halted?**
2. **Can the editing rates of endogenous sites be determined in parallel from RNA-seq data?**
3. **Are the underlying editing rates robust between replicates and do they correlate between cell types?**
4. **How many genes contain sites with robust editing rates?**
5. **Can the editing rates be used to identify when transcription was halted?**

2.2 Theoretical model of A-to-I editing

The lifecycle of a human mRNA as introduced in Chapter 1 Section 1.2 begins with transcription and ends with decay. For the purposes of modelling A-to-I editing here, it is assumed that:

1. The entire transcript is produced at a single time point, $t = 0$
2. The transcript is equally available for editing at all times between $t = 0$ and $t = t_\delta$ where t_δ is a random variable corresponding to the time at which the transcript is degraded
3. The entire transcript is degraded at a single time point $t = t_\delta$

Rodriques et al. (2020) found that the relationship between the fraction of edited transcripts containing site i, y_i, to the time t since those transcripts were synthesised can be described by

$$y_i(t) = 1 - e^{-\lambda_i t}, \tag{2.1}$$

where λ_i is a parameter denoting the rate at which the editing level increases. However, since this does not fix the model uniquely, an explicit link between the observed editing level at site i and the probability of that site being edited on a single transcript over time was needed. Furthermore, since transcripts contain multiple sites, a model that would link the editing states of multiple sites on a transcript to the age of the transcript was desired.

Let the state of site i be denoted by the discrete random variable $X_i \in \{0, 1\}$, where $X_i = 0$ is defined to be equivalent to the unedited base adenosine, A, at site i and $X_i = 1$ is equivalent to the edited base inosine, G (inosine appears as guanosine in sequencing data).

It is assumed that for N sites, all sites are initially unedited, so that at time zero $\boldsymbol{x}(t = 0) = \mathbf{0}$, where $\boldsymbol{x} \in \{0, 1\}^N$.

To model the time it takes for a site to be edited, let $T_i \in (0, \infty)$ be a random variable denoting the survival time of the adenosine at site i. Since it is assumed that editing at one site does not influence editing at any other site, the state X_i is assumed to be independent of all the other states. Thus, T_i is independent of the other survival times.

Based on the model of Rodriques et al. (2020) (Equation 2.1), T_i can be modelled as exponentially distributed with parameter λ_i. So, for $t \geq 0$,

$$P(T_i > t) = e^{-\lambda_i t} \tag{2.2}$$

and

$$P(T_i \leq t) = 1 - e^{-\lambda_i t}, \tag{2.3}$$

which is the cumulative distribution function (cdf) of the survival time of the adeno-

sine at site i.

The survival time of an individual adenosine at site i can be related to the expected editing level of site i over many reads through the linearity of expectation of independent random variables. Let M denote the total number of reads containing site i. The expectation for the fraction of reads on which site i is edited at time t is given by

$$\begin{aligned} E\Big[\frac{\text{num. edited reads}}{M} \text{ at time } t\Big] &= \frac{1}{M}\sum_{M \text{ trials}} E\big[X_i \text{ at time } t\big] = E\big[X_i \text{ at time } t\big] \\ &= 0.P\big(X_i = 0 \text{ at time } t\big) + 1.P\big(X_i = 1 \text{ at time } t\big) \\ &= P\big(T_i < t\big) = 1 - e^{-\lambda_i t} = y_i\,, \end{aligned} \tag{2.4}$$

where the first line invokes linearity of expectation and the fact that $X_i = 1$ at time t is equivalent to a survival time of less than t. Thus, the editing rate equation (Equation 2.1) is consistent with the expectation of a survival time model.

The above model assumes that all sites are unedited at time zero (recall that $\boldsymbol{x}(t = 0) = \mathbf{0}$). However, for the 'calibration experiment' detailed in Section 2.3, an alternative scenario is now considered where time, t, is relative to when ActD was added ($t = 0$). In this scenario, the fraction of edited reads at site i, at time t, is approximated by

$$y_i(t) = 1 - a_i e^{-\lambda_i t}. \tag{2.5}$$

where $1 - a_i \in [0, 1]$ is the fraction of edited reads at site i at time $t = 0$. The survival time T_i can be modelled in this scenario by

$$P(T_i \leq t) = (1 - a_i) + a_i(1 - e^{-\lambda_i t}),\ t \in (0, \infty), \tag{2.6}$$

This formulation is essential in the context of the calibration experiments described in this Chapter. Since the values of y_i are measured from when transcription is halted with ActD (which corresponds to the experiment time $t = 0$), the editing sites typically have a non-zero value for y_i, corresponding to the editing level immediately before ActD was added. a_i is used to reconcile the fact that experimental time $t = 0$ does not correspond to time since the transcripts were produced.

Thus a_i encodes the initial age at the moment that transcription is halted. Specifically,

$$y_i(0) = 1 - a_i \Rightarrow a_i = 1 - y_i(0), \tag{2.7}$$

which can also be expressed as

$$a_i = e^{-\lambda_i \tau_{0,i}}, \tag{2.8}$$

where $e^{-\lambda_i \tau_{0,i}}$ is the mean of $e^{-\lambda_i \tau^*_{0,i}}$, with $\tau^*_{0,i}$ denoting the random variable that is the age of the reads containing site i at the time ActD is added ($t = 0$). Thus, an estimate for the age of the reads containing site i can be related to the time since ActD was added by

$$\bar{\tau}_i = t + \tau_{0,i} \tag{2.9}$$

and substituting a_i using Equation 2.8 and t using Equation 2.9 gives

$$y_i(t) = 1 - e^{-\lambda_i (t+\tau_{0,i})} = 1 - e^{-\lambda_i \bar{\tau}_i}. \tag{2.10}$$

Finally, solving for $\bar{\tau}_i$ gives an estimate of the age of the reads containing site i in terms of the editing level, y_i and the editing rate, λ_i:

$$\bar{\tau}_i = -\frac{\ln(1 - y_i)}{\lambda_i}. \tag{2.11}$$

Thus, Equation 2.5 and the parameter a_i are only used for determining λ_i for site i in the context of the calibration experiment.

The site age estimates, $\bar{\tau}_i$, for the set of all sites annotated to a gene, $i \in I$ are assumed to be estimating the mean age of the same population of transcripts and thus can be averaged to improve the precision of the estimate of the mean population age using the central limit theorem. Thus, the mean age of the population of transcripts - or simply: the 'mean gene age', $\bar{\tau}_g$ - is given by

$$\bar{\tau}_{\mathrm{g}} = \frac{1}{|I|} \sum_{i \in I} \bar{\tau}_i. \tag{2.12}$$

This background is sufficient for the results contained in this Chapter, which is concerned with how the observed changes in editing level after transcriptional halting relate to the process of editing at individual sites. In Chapter 3, the editing of multiple sites on a single transcript is used to infer the age of the transcript. Since the theoretical framework for that follows logically from the above, it is explained here:

For a single transcript containing N sites, the likelihood is defined to be the probability of the system of N sites being in the state $\boldsymbol{x}$ for a transcript (i.e. read) of age τ_r

$$L(\tau_r|\boldsymbol{x}) = \prod_{i \in G} P(T_i < \tau) \prod_{j \in A} P(T_j > \tau) = \prod_{i \in G} (1 - e^{-\lambda_i \tau}) \prod_{i \in A} e^{-\lambda_i \tau}, \tag{2.13}$$

where G is the set of edited sites and A is the set of unedited sites. The value of the parameter τ_r is then estimated with maximum likelihood estimation (MLE) (as described in Methods Section 6.3.2) to give the estimated transcript age, $\hat{\tau}_r$, which has a range between 0 hours (when all sites are unedited e.g. $\mathbf{x} = \mathbf{0}$) and ∞ (when all sites are edited, e.g. $\mathbf{x} = \mathbf{1}$).

2.3 Measurement of A-to-I editing rates in hiPSC-derived neurons

Based on the work of Rodriques et al. (2020) (as described in Chapter 1, Section 1.5) it was hypothesised that endogenous A-to-I editing sites may exhibit robust editing over time and that these rates could be determined using a similar method to their 'calibration' protocol. This protocol (described in the Appendix, Section 6.7) involves halting transcription with Actinomycin D (ActD), taking samples at time points after ActD is added and then fitting a model to determine the editing rates. In total, four calibration experiments were conducted on human cell types, which are summarised in Table 2.1.

	Cortex	**Midbrain**	**NLambda HEK**	**Endogenous HEK**
Time points in calibration (h)	0,2,4,6,8	0,2,4,6,8	0,1,2,3,4,5,6,7,8,9,10,11,12,16,24,38	0,4,6,8,12,24
Significant sites (number of genes)	1,044,970 (13,171)	42,684 (3,523)	428,418 (16,168)	157,774 (8,007)
Fit sites (number of genes)	162,741 (7,811)	30,154 (2,851)	68,599 (6,386)	60,952 (3,459)
Fit sites in 5'UTR (%)	0.43%	0.88%	1.45%	0.57%
Fit sites in exons (%)	5.29%	10.6%	13.7%	8.30%
Fit sites in introns (%)	82.2%	58.0%	42.9%	46.5%
Fit sites in 3'UTR (%)	12.1%	31.0%	42.0%	44.6%

Table 2.1: **Summary of calibration experiments performed in different cell lines.** The cell types are labelled as follows: 'Cortex' and 'Midbrain' denote human iPSC-derived cortical and midbrain neurons, respectively. 'NLambda HEK' denotes HEK293 cells that had been transfected with an modified ADAR2 construct, ADAR2(E448Q)-NLambda (See Section 2.4). 'Endogenous HEK' denotes HEK293 cells that were not transfected with an editing construct and thus contained only their endogenous ADAR proteins. 'Time points in calibration (h)' denotes when samples were taken and sequenced relative to the time at which transcription was halted with Actinomycin D (ActD). 'Significant sites' are those that have a statistically significant increase in editing level 8 hours after ActD addition. Significance is determined using the JACUSA2 software (Piechotta et al., 2021) where the significance level is a Z score greater than 1.96. 'Fit sites' are a subset of 'Significant sites' that are fit by Equation 2.5 and pass the filtering criteria described in Section 2.3. The bottom 4 rows specify the percentage of sites annotated to different regions of the genes. UTR stands for untranslated region.

The initial calibration experiments were performed in human induced pluripotent stem cell (iPSC)-derived cortical and midbrain neurons. These were chosen due to the higher reported editing in the central nervous system relative to other cell types (Ramaswami and Li, 2014).

Short-read RNA sequencing was performed on samples taken at 0h, 2h, 4h, 6h and 8 post-ActD treatment. To determine if edits were accumulating over time, JACUSA2 (Piechotta et al., 2021) was used to determine differentially edited sites between the 8-hour and 0-hour post-ActD conditions. In the cortical neurons, 1,044,970 sites were identified as significantly increasing in editing level across 13,171 genes (Z score greater 1.96 from JACUSA2's likelihood ratio test). These sites constitute the list of 'significant sites' ('sig sites' for short, Figure 2.1A).

Having determined that many sites accumulated edits over time, it was hypothesised that some of the 'sig sites' may accumulate edits at a robust rate. Using a non-linear least squares algorithm (Methods Section 6.2.1), Equation 2.5 was fit to each site i to determine the editing rate, λ_i (Figure 2.1B). The results were filtered as follows to yield the list of 162,741 'fit sites' across 7,811 genes (Figure 2.1A):

- Site i has an editing rate (λ_i) between $0.01h^{-1}$ and $2h^{-1}$
- Site i has a coefficient of correlation (R^2) for the fit of Equation 2.5 to the calibration data of greater than 0.4
- Site i has significant editing (as determined by JACUSA2) in at least four time points from the calibration experiment
- Site i has a mean age, $\bar{\tau}_i$, of less than 25 hours prior to ActD addition.

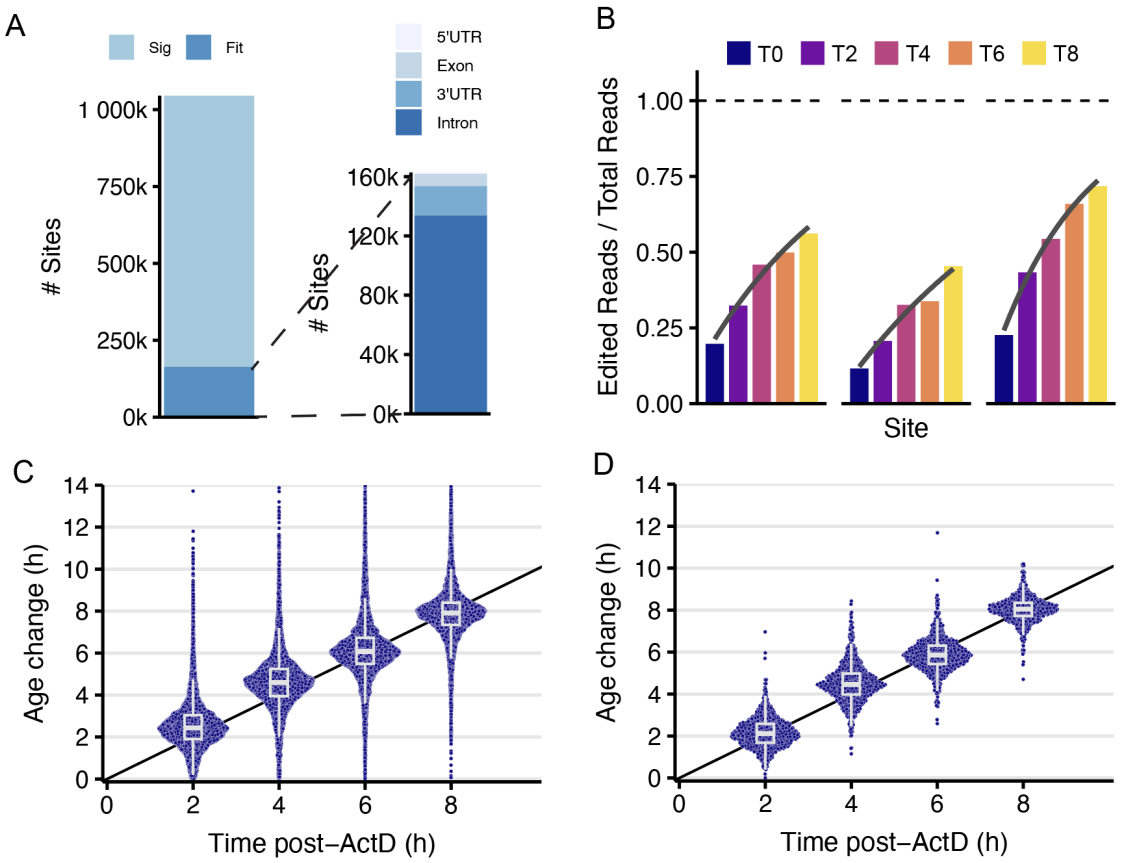

Figure 2.1: **Prediction of transcript ages from endogenous ADAR-mediated A-to-I editing in human cortical neuron culture.** (A) Left: 1,044,970 adenosines are significantly differentially edited in hiPSC-derived cortical neurons 8 hours after adding Actinomycin D. Right: Of those editing sites, 162,741 are well fit by the calibration equation (Equation 2.5) The regions that sites are annotated as belonging to within genes are shown as stacked bars. (B) Editing fractions at three sites in the 3'UTR of the EIF3M gene are shown at 2 hour increasing time intervals following ActD treatment. The fit of an exponential CDF is shown in grey. (C) The mean age of each gene (purple) is calculated from the short-read HEK calibration data and the time change relative to the 0 h time point is shown. A black line of gradient 1 is shown as a guide. Boxplots show the 25$^{\text{th}}$, 50$^{\text{th}}$ and 75$^{\text{th}}$ quantiles, whiskers extend to the value at most 1.5 × IQR from the respective hinge. ($n = 7,811$ genes) (D) Same as C but with a subset of fit sites that appear in the 3'UTR of genes with at least five 3'UTR sites ($n = 773$ genes).

The same procedure was performed for iPSC-derived midbrain neurons, which resulted in 42,684 sig sites ($n = 3,523$ genes) and 30,154 fit sites ($n = 2,851$ genes). For both the cortical and the midbrain neurons, most of the fit sites were found to be in either introns or in the 3'UTR region of genes (Figure 2.2A,B). Of note, the cortical neurons exhibited a much larger intronic fraction than the midbrain neurons (82.2% vs 58.0%).

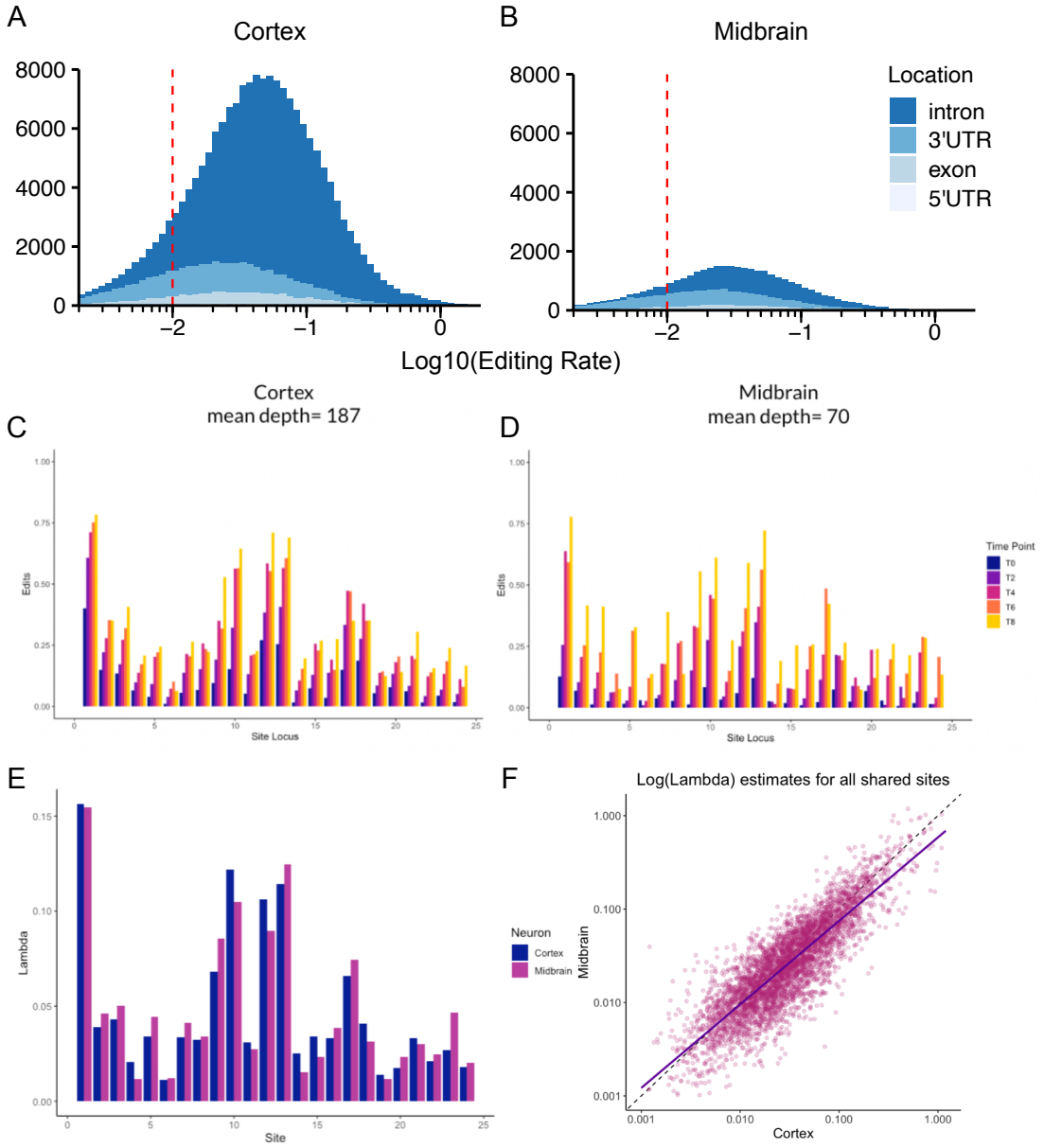

Figure 2.2: **Comparison of calibration experiments performed in iPSC-derived cortical and midbrain neurons.** (A and B) Stacked histograms showing the locations of fit sites over the range of log-transformed editing rates. Red dotted line indicates the lower boundary of editing rate for inclusion as a fit site. (C and D) The accumulation of edits post-ActD is shown at each fit site in the gene PPM1K. (E) The editing rates (λ) from cortex and midbrain are shown side-by-side for each site. (F) Log-transformed editing rates of the two datasets are plotted against each other, with a linear model plotted as a purple line.

Comparing the editing rates at shared sites between the two datasets revealed a strong correlation between the two neuronal cell types (Figure 2.2C-F), despite the large difference in sequencing depth between the two experiments (Spearman corre-

lation coefficient = 0.84).

With evidence that A-to-I editing was a robust process across cell types and Equation 2.5 which provides a link between editing rate (λ_i), editing level (y_i) and time (t), I reasoned that rearranging Equation 2.5 would estimate the age of the pool of transcripts containing site i. In this context, the 'age' is defined as the mean duration of time that has passed since transcription, i.e. how long the individual transcripts have existed in the cell. The mean age was calculated for each site using Equation 2.11.

For sites $i \in I$, the set $I \subset \mathbb{Z}^+$ contains the fit sites annotated to a gene, g, which are assumed to be transcribed together and (under the assumption of independence of editing) each provide an independent estimate of the mean age. $\bar{\tau}_g$, of the pool of transcripts (referred to as the 'gene' for simplicity). Whilst it is both possible and likely that any given gene can produce multiple differentially regulated isoforms that also have different editing profiles, resolving those differences with short-read sequencing is impossible since the reads are too short for the splicing pattern to be determined. Thus - whilst being a reductive assumption - isoforms are ignored.

Under this assumption, the mean age of a 'gene' is given by:

$$\bar{\tau}_{\mathrm{g}} = \frac{1}{|I|} \sum_{i \in I} \bar{\tau}_i, \tag{2.11}$$

where $|I|$ is the number of fit sites annotated to the gene.

For the cortex calibration dataset, the mean gene age was calculated for every gene with at least one fit site ($n = 7,811$ genes) using Equation 2.11. Since the duration of time between sequencing steps in the calibration experiment is known, the performance of the mean age model was tested by its ability to correctly predict the change in time over the course of the calibration experiment. For each gene at each time point, the mean age was expressed as the change relative to the 0h time point

in the experiment as

$$\Delta\bar{\tau}_g = \bar{\tau}_{g,t} - \bar{\tau}_{g,t=0}, \tag{2.14}$$

and plotted in Figure 2.1C. To quantify the accuracy of the predictions, a linear model was fit to $\Delta\bar{\tau}_g$ for each gene over the time points from calibration time point (2h, 4h, 6h, 8h - the 0h was not present the mean ages are expressed relative to it) (Methods Section 6.2.1). The gradient from the linear model, $\hat{\beta}_g$ and the R^2 value was determined for each gene, which had mean values over all the genes of 0.902 and 0.826, respectively. These results showed that the predicted age increases correlated strongly with the known time change between calibration time point, demonstrating that endogenous A-to-I editing encodes the passing of time.

To further understand the source of variance in the accuracy of the age measurements, a subset of fit sites that only included 3'UTR sites from genes that had at least five fit sites in the 3'UTR was created ($n = 18,505$ sites). The selection of 3'UTR sites was based on the rationale that these sites were more likely to be present in mature mRNA than the intronic sites, and that they were more likely to be found in *Alu* repeats than sites annotated to other parts of the gene body, thereby selecting sites for which there is a well characterised mechanism for editing by ADAR1. Comparing this subset of 18,505 sites with the full list of fit sites showed an improvement of the linear model fit (the mean gradient increased to 0.949, the mean R^2 increased to 0.93) demonstrating that it was an effective heuristic for selecting high-performance sites (Figure 2.1D).

2.4 Measurement of endogenous editing in HEK cell culture

Encouraged by the agreement of the results between the two neuronal calibrations, a further calibration experiment was carried out in HEK293 cells. Since it had been previously reported that HEK cells had lower editing rates than primary cells

or iPSC-derived cell lines (Schaffer et al., 2020), the HEK cells were transfected with a construct that would express an engineered ADAR protein. My collaborator Ali Ghareeb designed and screened several ADAR2 variants and fusion constructs. Based on it having the greatest number of detected editing sites amongst the screened variants, the construct ADAR2(E448Q)-NLambda was selected. This protein comprises an ADAR2 variant with the hyperactive mutation E44Q, and the fused anti-terminator protein N from λ bacteriophage (used by others to tether proteins to RNA (Baron-Benhamou et al., 2004; Montiel-Gonźalez et al., 2016)).

Ali performed a calibration protocol with HEK293 (HEK, for short) cells expressing this NLambda construct as described in the Appendix Section 6.7 (the time points are given in Table 2.1). The samples were sequenced on a NovaSeq (Illumina Technologies) with an average of 225 million reads per time point. This dataset is referred to this as the 'NLambda calibration'.

The editing rates were calculated as described above for the neuronal calibrations.

Inspection of the editing levels at individual sites revealed that the fraction of editing often plateaued below saturation. To account for this effect, the equation fit to the neuronal calibration data (Equation 2.5 was modified to

$$y_i(t) = \beta_i(1 - a_i e^{-\lambda_i t}), \tag{2.15}$$

with the mean age $\bar{\tau}_i$ of site i given by

$$\bar{\tau}_i = -\frac{\ln(1 - \frac{y_i}{\beta_i})}{\lambda_i}, \tag{2.16}$$

where the new parameter, β_i, was fit for each site i. Across the dataset, β ranged from 0.09 to 1.0 with a mean of 0.69 and standard deviation of 0.279 ($n = 53,341$ fit sites) (Figure 2.3A). However, using β parameters in age estimation proved difficult since, in many time points, the observed y_i value was greater than β_i - in which

case equation 2.16 is undefined. This was the case for a mean of 5.2% of sites in the 0h to 12h time points from the calibration experiment and increased to a mean of 14.3% for time points between 10h and 36h. Furthermore, when the β and λ parameters determined from the NLambda calibration were used to calculate ages in untreated HEK cells from a separate experiment that my collaborators performed, the incidence increased to 58.6%. Thus, the β parameter was abandoned in favour of an alternative solution.

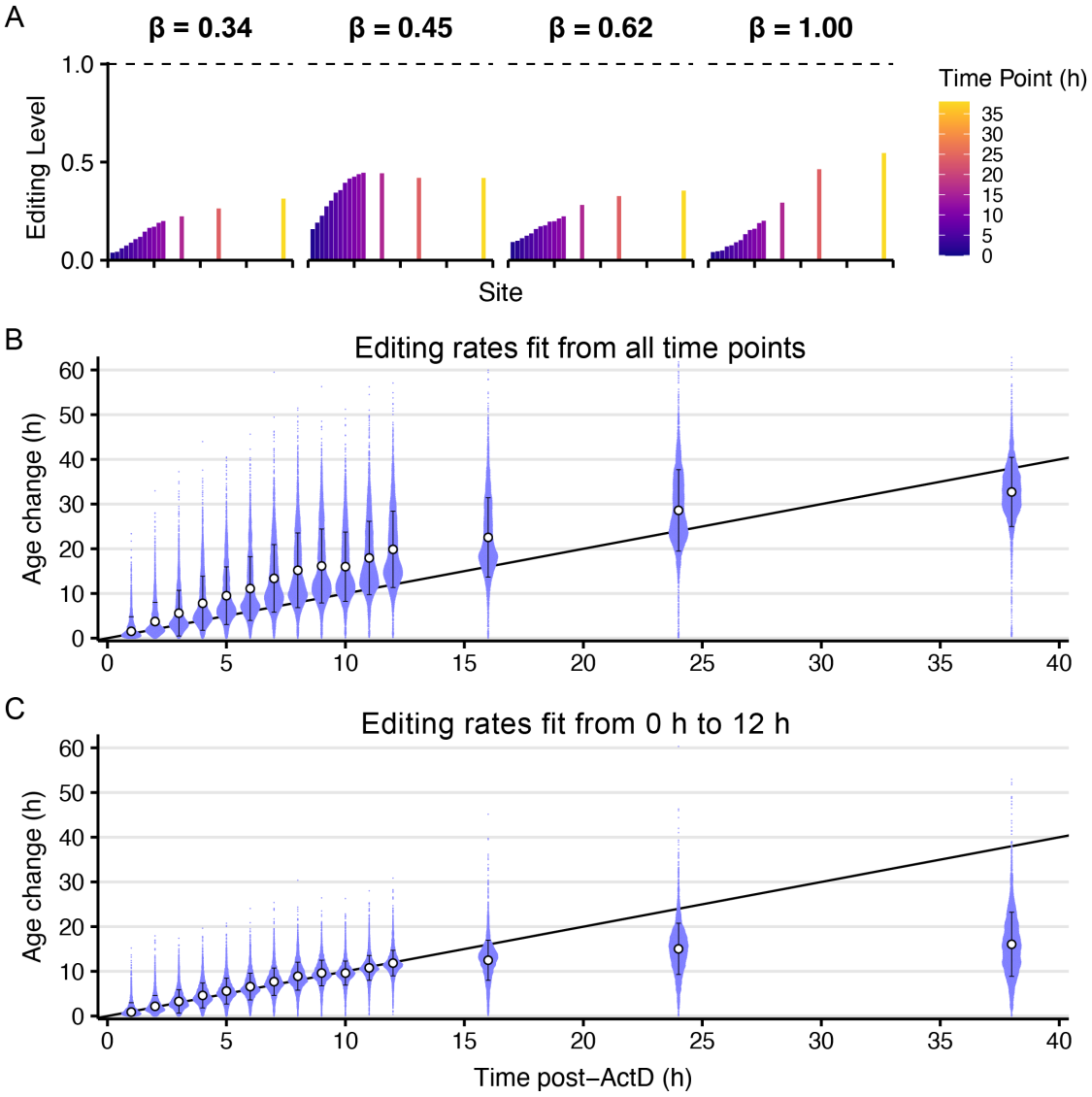

Figure 2.3: **Some editing sites do not tend towards being fully edited with increasing time** (A) Four sites from highly-expressed transcripts taken at random from different bins of β. The accumulation of edits during the HEK293 calibration experiment is shown, with the fit β value shown at the top. (B and C) The change in predicted age relative to the 0 h time point is calculated for each gene ID for each time point and plotted as blue points as a 'beeswarm' distribution. The predicted ages are generate using λ parameters fit from all time points in the calibration (B) or only up to and including the 12 h time point (C). The mean (white circle) and standard deviations (black lines) are overlaid. The black diagonal trace denotes the line $y = x$.

Since it has been reported that ActD can cause cellular stress, it was reasoned that the apparent under-editing at late time points may be due to prolonged exposure to ActD causing severe disruption to cellular processes beyond simply halting transcription (Lugowski et al., 2018). As such, fitting to only a subset of the calibration

time points was explored.

Comparing the predicted mean age change vs the actual known time change between time points in the calibration experiment showed that fitting to all 16 time points introduced a bias for over-estimating the age of genes in the majority of time points (Figure 2.3B). Therefore, time points later than 12 hours post-ActD addition were removed and the Equation 2.5 fit to the truncated dataset (Figure 2.3C). This change decreased the standard deviation of the age predictions from 12.32 h to 4.47 h for time points up to 12 hours. Since the typical half-life of a human mRNA is thought to be between 5 to 10 hours, improved performance on predicting mean ages in this range was prioritised over performance on older ages. Thus, determining the editing rates from calibration time points no later than 12 hours post-ActD effectively mitigated the issue of lower than expected editing at extreme time points.

Calculating the editing rates on this portion of the NLambda calibration data yielded 428,148 sites with significantly increased editing 8 hours post ActD and 68,599 fit sites (Table 2.1, Figure 2.4A). These fit sites had a mean editing rate of 0.058 per hour, corresponding to an expected editing time of 17.2 hours.

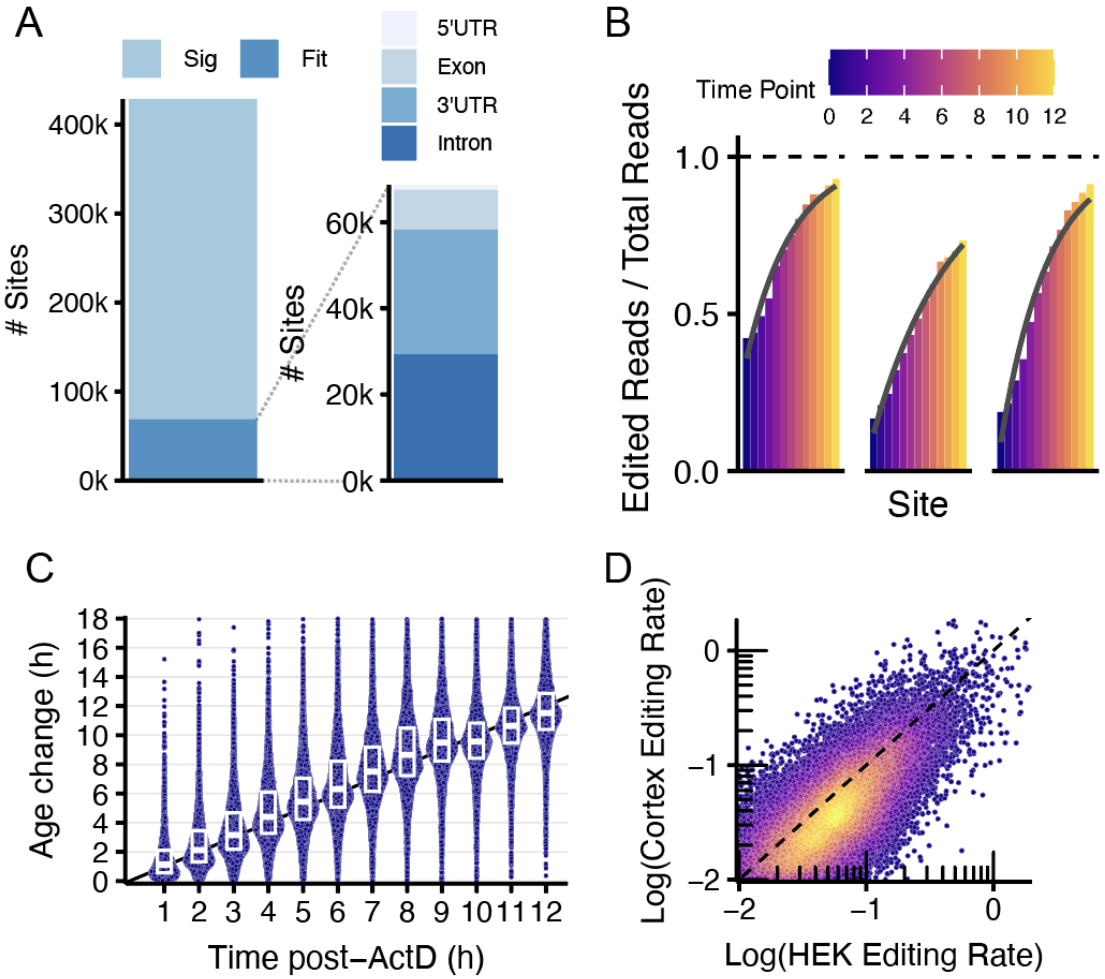

Figure 2.4: **HEK calibration with NLambda construct** (A) Left: 428,148 adenosines are significantly differentially edited in HEK cells expressing an ADAR2(E488Q)-Nlambda plasmid 8 hours after adding actinomycin D. Right: Of those editing sites, 68,599 are well fit by an exponential CDF, and their location within genes are shown as stacked bars. (B) Editing fractions at three sites in the 3'UTR of the GATC gene are shown at monotonically increasing time intervals following ActD treatment. The fit of an exponential CDF is shown in grey.(C) The mean age of each gene is calculated from the short-read HEK calibration data and the time change compared to the T0 time point (plotted as purple points). A black line of gradient 1 is shown as a guide. Boxplots show the 25th, 50th and 75th quantiles, whiskers extend to the value at most 1.5 × IQR from the respective hinge. (D) The correlation of editing rates at shared sites in the ADAR2(E488Q)-Nlambda expressing HEK293 calibration and hiHPSC-derived cortical neuron calibrations are shown as points coloured by 2D kernel density and plotted on log10-log10 axes

The performance of the mean age model using the editing rates from the NLambda calibration was determined using the same $\Delta\bar{\tau}_g$ analysis as for the cortical neurons in Figure 2.1. The mean gradient from the linear regression to each gene was 0.98 with a mean R^2 of 0.75 (Figure 2.4C). Thus the editing rates from the NLambda

dataset more closely agreed with the known change in time between calibration time points than the cortex dataset, yet there was more unexplained variance across the set of genes as evidenced by the lower R^2 value.

The correlation of the editing rates for shared sites between the NLambda with the cortex dataset was determined using Pearson's method which produced correlation coefficient of 0.69 (Figure 2.4D), indicating a good agreement of the editing rates between the cell types (even with NLambda expressing the engineered editor construct).

The editing rates were also determined in HEK cells without the NLambda construct by performing a separate calibration experiment in HEK cells expressing endogenous ADAR only (Table 2.1). Comparing the distributions of the editing rates derived from the HEK and NLambda HEK calibration experiments showed that the primary effect of NLambda was to introduce a large number of slowly-edited sites ($\lambda < 0.01\text{h}^{-1}$) (Figure 2.5A,B). An editing rate of 0.01h^{-1} corresponds to an expected time of editing of 100 hours, which is far in excess of the half-lives of most human mRNA. Therefore, the majority of the sites introduced by NLambda cannot be used for age inference.

The location of the editing sites was compared between the NLambda and the endogenous HEK calibration experiments. This revealed a large increase in the proportion of editing sites found in exons in the NLambda dataset compared with the other datasets that did not express the construct (for 8.3% for the endogenous HEK, 13.7% for the NLambda HEK) (Table 2.1, Figure 2.5A,B).

Despite the differences between the NLambda and HEK calibration datasets, using the NLambda dataset for downstream analysis was appealing since it was sequenced at higher depth and more time points were taken. However, in order to use the NLambda data, two key concerns had to be mitigated. First, whilst sites that are introduced by NLambda should not appear in samples not expressing the construct, if editing *is* observed at those sites, then using the editing rate from the NLambda

calibration will introduce noise to the age estimates. Second, if the presence of the NLambda construct alters the editing rate at a site that is also edited by endogenous ADAR, then age estimates based off that site will be inaccurate (i.e. if the editing rates of a 'shared site' between the two calibration datasets differ).

To investigate whether the editing rates at sites edited by both endogenous ADAR and NLambda were similar, I compared the correlation between the two and found a strong agreement (Pearson's correlation coefficient = 0.88) - although the NLambda editing rates were slightly lower across the set (Figure 2.5C). The analysis was broadened to not measure the change not just at shared sites but also at sites only edited by NLambda. This was done by applying the editing rates from the NLambda calibration to all of the editing data from either the endogenous calibration or the NLambda calibration (using the 0h time point - i.e. steady state - from each) (Figure 2.5D). Only 30 out of the 1,686 genes for which the mean age could be calculated had a statistically significant difference in age between the two datasets (Student's t-test, Benjamini-Hochberg corrected). Together these analyses suggested that editing levels from endogenous ADAR1 was sufficient in HEK cells for the cell line to be used for further method development and that editing rates from the NLambda calibration dataset (despite having introduced editing sites not targeted by endogenous ADARs) introduced only a small bias and could still be used in downstream analysis.

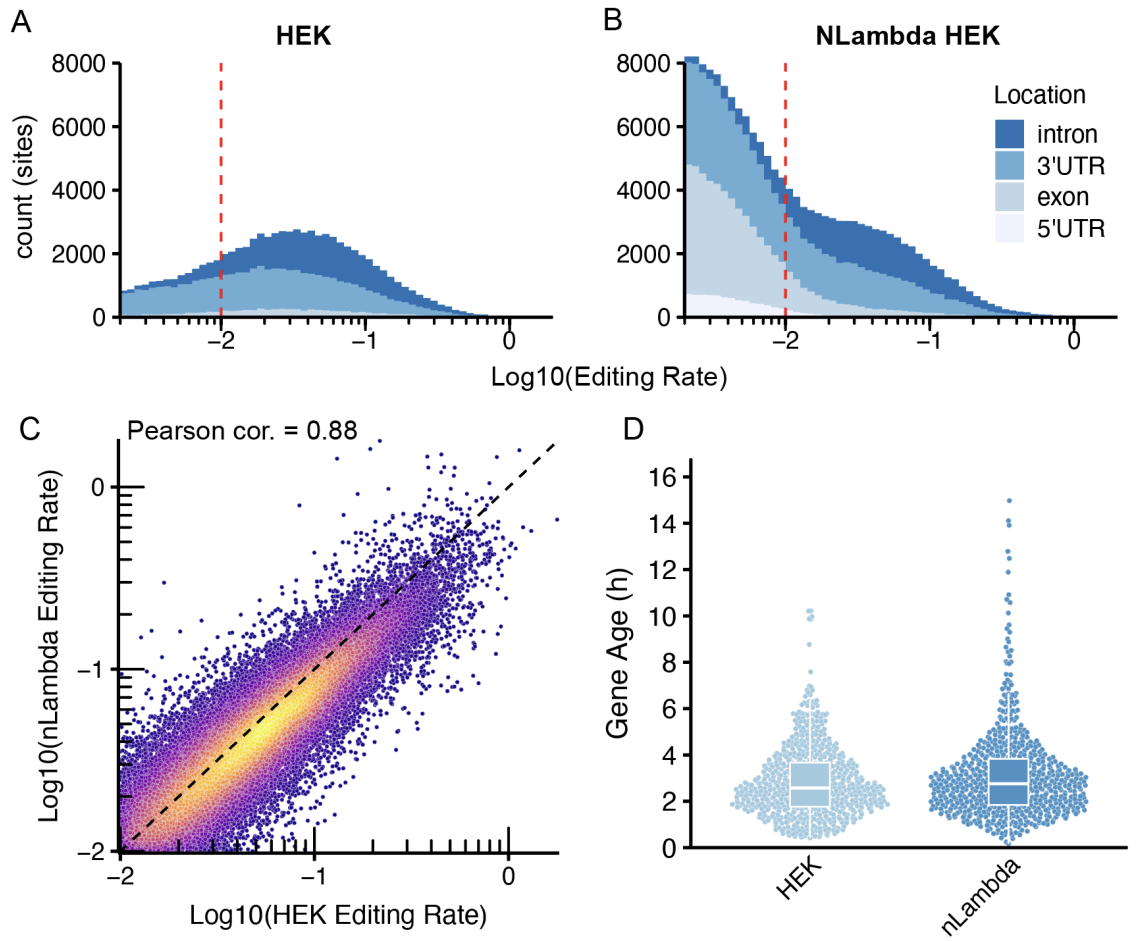

Figure 2.5: **The NLambda construct primarily introduces sites with slow editing rates.** (A and B) Stacked histograms showing the locations of fit sites over the range of log-transformed editing rates. Red dotted line indicates the lower boundary of editing rate for inclusion as a fit site. (C) The log-transformed editing rates of the shared fit sites between the NLambda and HEK calibration datasets are shown as points coloured by 2D kernel density. (D) The predicted mean ages for the transcripts from shared genes (points) in the 0 h time point from the respective calibration experiments. Boxplots show the 25th, 50th and 75th quantiles, whiskers extend to the value at most 1.5 × IQR from the respective hinge.

Having established that editing rates broadly agreed between NLambda, endogenous HEK, Cortex and Midbrain it was reasoned that it may be possible to combine the calibration datasets. The motivation for this is that each of the cell types expresses a unique subset of the genome and so combining them would enable inference on a wider range of genes. To this end, the NLambda calibration data and the Cortex calibratinod data were combined (as detailed in Methods Section 6.2.1). This yielded a set of 201,223 fit sites across 8,685 genes.

2.5 Discussion

The overarching research questions of this chapter were whether the rates of endogenous A-to-I editing could be measured in a high-throughput experiment and whether they encoded the passage of time. By adapting the calibration protocol of Rodriques et al. (2020), the editing rates at hundreds of thousands of endogenous editing sites were measured, for the first time, in three human cell types. Analysis revealed a large range of editing rates and detection of different sets of editing sites in different experiments. However, for editing sites shared between experiments, the editing rates correlated well and, importantly, were able to accurately predict the amount of time that had passed since transcription was halted in the calibration experiment. Thus, both of the two primary research questions were answered.

As shown in Figure 2.1C & D, the way that the list of 'fit sites' is defined has implications for the results obtained. Whilst it was found that subsetting only to sites in the 3'UTR improved the performance of the age model, it came at the expense of a 10-fold reduction in the number of genes that were measured (from 7,811 to 711 genes). The method used to filter sites was based on a heuristic approach and was by no means optimised. There are almost certainly better ways to filter the editing sites that arrive at a compromise between accuracy and applicability - such as by constructing more robust error metrics for the error of determining λ_i from the calibration data, and for the error of the mean gene age value (which is done in the next Chapter). However, the filtering thresholds applied here are used throughout the following chapters, and the impact of the specific values chosen on the results is not investigated. To an extent, this is similar to much of the A-to-I editing literature, where arbitrary minimum thresholds of the number of reads and/or the number of observed edits are common (Bazak et al., 2014; Ramaswami and Li, 2014; Piechotta et al., 2021). Whilst those thresholds, along with the sequencing depth used in different experiments, will affect how many editing sites are discovered, the consequences of these thresholds to the statistical significance of the results are

rarely detailed in full.

Drawing a distinction between the observed degree of editing at a site and the underlying (and seemingly intrinsic) editing rate is important for several reasons. The first is in relation to the concept of steady state transcription. If the turnover of a the mRNA of a specific gene is far from steady state - for instance, if there is a rapid increase in the transcription rate - then the observed editing may be temporarily altered. From standard editing analysis that simply compares the editing levels between the two conditions, one might conclude that the observed change marks a regulated shift to a different steady state. In reality, however, the parameters underlying the editing process - that is: the affinity of ADAR1 for binding the RNA species, ADAR1's catalytic activity and whether the availability of ADAR proteins is limiting - are likely unchanged (given ADAR1's broad set of substrates, the increase in new transcripts would have to substantially increase the total number of ADAR1-binding transcripts in the cell for the availability of ADAR1 proteins to be substantially altered). The second reason is that if the underlying editing rates are robust across experiments, cell types and perturbations, then the editing process may be used as a molecular clock to measure how long individual transcripts have existed in the cell. I have already introduced this concept in this chapter, where I used the predicted difference in age between calibration time points vs the actual change as a validation method (Figures 2.1, 2.4, 2.3, 2.5).

An outstanding question from this Chapter is whether sites that have significant editing but are not well fit by the editing rate model (Equation 2.5) are edited by a different process. Such an analysis was not performed here but the existence of other processes that might affect the editing rate is conceivable and there are several unexplained phenomena in the calibration data. For instance, the results of including the β_i parameter in the age modelling (reported in Section 2.4) suggested that the phenomenon of the editing rate seemingly decreasing after many hours (> 12h) of ActD exposure was site-specific. It was also not determined whether this effect was due to ActD specifically. Conversely, early time points in the calibration

experiment seemed to display fewer edits than expected at some sites (for example, the first site in Figure 2.4B). This may also be an artefact of ActD: specifically, that there is a time delay between when it is added to the cell culture and when it has diffused into cells and halted transcription. A similar effect has been considered by others for metabolic labelling experiments (Rummel et al., 2023), where modelling a diffusion time of 4sU substantially improved the model performance (a procedure they term 'recalibration').

In addition to potential side-effects of ActD there are biological effects that may affect how the editing rate is calculated. For instance, it was not determined which editing sites were targeted by the p150 isoform of ADAR1, p110 or both. Given that the p110 isoform is mostly nuclear and p150 mostly cytosolic, this offers a potential explanation for why some sites are not well fit by the editing rate equation (Equation 2.5) yet do have significant editing. If those sites are only edited by p110 and the transcripts are exported from the nucleus quickly (on the order of a few hours) then they may not be editable for the remainder of the calibration experiment. However, Sun et al. (2021) did not find any p110-only editing sites in their analysis and so it seems unlikely that they account for such a large proportion of the editing sites here. Instead, the drop-off from the number of significant sites to fit sites may instead be an issue of sensitivity, where the editing rates at these sites is so slow that an even greater sequencing depth would be required to determine the editing rate. However, an extremely high sequencing depth was used to sequence the cortical neuron calibration here and yet the only 15.6% of the sig sites were well fit. In summary, further experiments may be required to identify why so many sites are not well fit by Equation 2.5.

In the next Chapter, the model of mean gene age presented here is further developed and a new model for single-transcript age estimation presented.

Chapter 3

Modelling RNA age

"If I could save time in a bottle... "

– Jim Croce, Time in a Bottle (1972)

3.1 Introduction

The ability to order and regulate biological processes over time is a fundamental hallmark of cell biology. RNAs - and in particular messenger RNAs (mRNA) - serve as a critical and highly dynamic species that have the dual advantage of being information rich and facile to measure with modern RNA-sequencing technologies. In any given cell, the pool of mRNAs is in constant flux, shaped by ongoing synthesis via RNA polymerase II and decay through a range of degradation mechanisms (introduced in Section 1.2). Despite the constant turnover of mRNA in the cell, existing RNA-seq methods only provide a snapshot of the relative abundances of the mRNA of different genes. In recent years, many new methods have been introduced to probe expression dynamics but no single method provides a complete picture. Metabolic-labelling can measure the transcription and degradation kinetics of mRNA (Herzog et al., 2017; Schofield et al., 2018; Jürges et al., 2018; Rummel et al., 2023); RNA velocity methods estimate whether expression is increasing or decreasing by modelling the ratio of spliced and unspliced reads in single cell data (La Manno et al., 2018; Bergen et al., 2020); molecular recorders order transcriptional events over long-time frames but require extensive engineering and are low throughput interest (Bhattarai-Kline et al., 2022); and Live-seq (Chen et al., 2022) allows repeated measurements of *in vitro* cells but has limited yield. While powerful,

each of these methods offers only a partial view, and no single technique provides a unified, label-free, and gene-agnostic estimate of RNA age.

Rodriques et al. (2020) demonstrated the principle that A-to-I editing encodes the time since transcription on the transcripts themselves (in this case, on a specific 3'UTR RNA sequence). However, their system required both a fusion protein transgene and insertion of an artificial editing cassette into the 3'UTR of specific genes - restricting its use to *in vitro* models and a limited number of targets.

The previous chapter showed that the endogenous A-to-I editing rates at hundreds of thousands of adenosines in the human transcriptome can be measured and are sufficient to extend the principle of 'editing encoding time' to the endogenous setting. This was done through the concept of 'mean gene age' - the average age of transcripts from a given gene, which was calculated at each editable site (referred to as 'per-site age') and then averaged over all the sites for a gene. This Chapter begins by characterising the performance of this method on the various calibration datasets introduced in Chapter 2.

While per-site endogenous RNA age (ERA) provides accurate mean age estimates, it cannot resolve the full distribution of transcript ages in a population. To enable inference of individual transcript ages and gain temporal resolution beyond the mean, I introduce a new per-transcript age method: per-transcript ERA. This approach, utilising long-read sequencing and a likelihood model for the age of a single transcript from its editing pattern, enables estimation of single-molecule ages from endogenous processes alone. It requires no genetic engineering and can be applied to any long-read dataset from human cells with sufficient ADAR2 activity and sequencing depth.

The development of these methods in this Chapter was framed by the following research questions, which I return to in the discussion at the end:

1. **Can the mean age of a population of transcripts be estimated from the observed editing levels at sites with known editing rates?**

2. **Can the age of single RNA molecules be inferred from long-read sequencing data?**

3. **How do different sequencing modalities and data processing software affect RNA age estimates?**

4. **Can RNA age estimates be used to infer mRNA half-lives, and do these estimates agree with those from other methods?**

A schematic of the per-site ERA and per-transcript ERA is presented in Figure 3.1.

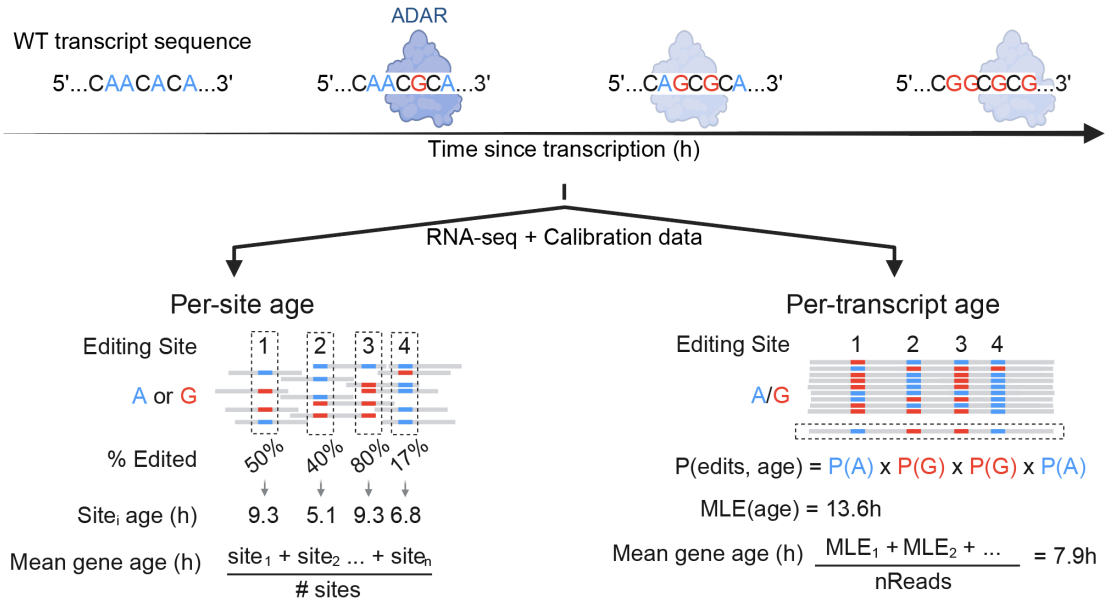

Figure 3.1: **Graphical overview of the per-site and per-transcript age methods.** ADAR1 introduces A-to-I edits that accumulate over time and can be used in the per-site age method (on either short or long-reads) or the per-transcript method (long-read only). Grey bars depict individual reads, blue and red marks indicate edited (red) or unedited (blue) sites. MLE stands for the maximum likelihood estimate of the age of single transcripts. (Created with biorender.com)

3.2 Per-site Age

As introduced in section 2.2 the amount of editing at a site can be used to estimate the mean age, $\bar{\tau}_i$ of the transcripts that contain that site by equation 3.1.

$$\bar{\tau}_i = -\frac{\ln(1 - y_i)}{\lambda_i} \tag{3.1}$$

where λ_i is the editing rate of site i as determined from a relevant calibration experiment and y_i is the observed fraction of edited reads over total reads (between 0 and 1). For estimating the mean age of a group of transcripts (e.g. all those arising from the transcription of the same gene (ignoring isoforms)), assuming each site provides an independent estimate of the mean age of the underlying transcript population, then - by the central limit theorem - the average of the per-site age estimates will approximate a normally distributed estimate of the group's true mean age (Figure 3.2A,B). In practice this is done by taking the arithmetic mean over all the using

$$\bar{\tau}_g = \frac{1}{|I|} \sum_{i \in I} \frac{-\ln(1 - y_i)}{\lambda_i}, \tag{3.2}$$

where I is the set of sites for a given gene, g (as already detailed in Chapter 2 Section 2.2).

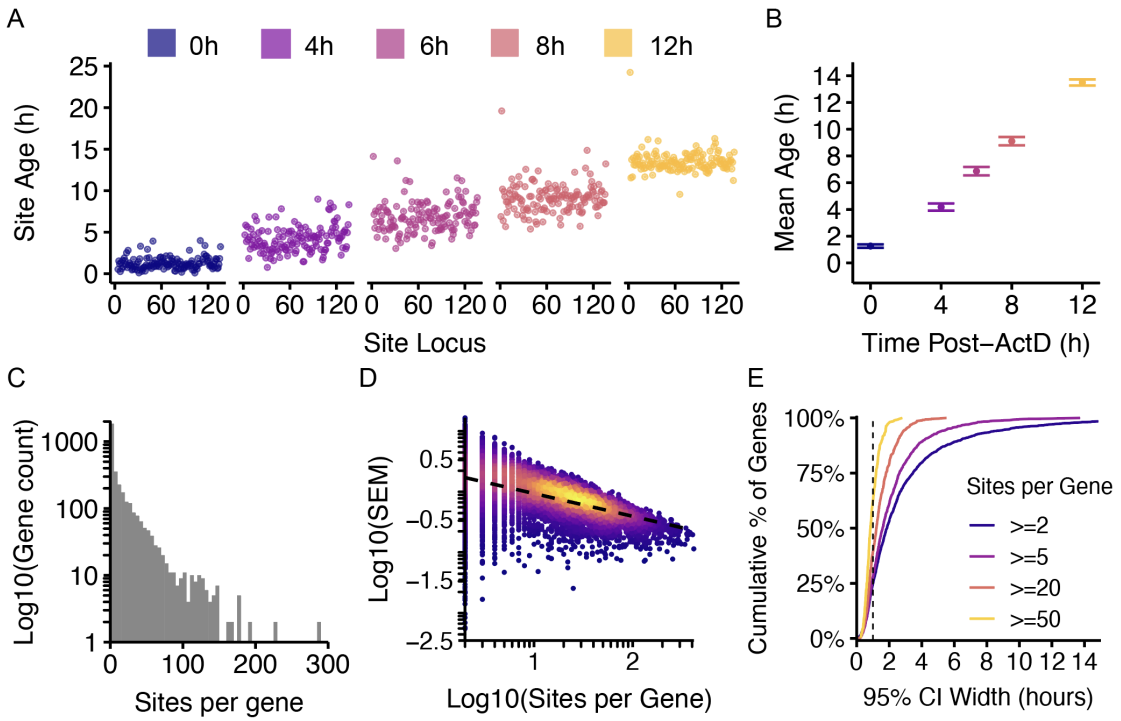

Figure 3.2: **Mean ages of genes can be precisely determined from per-site ages.** (A) The ages of individual sites (points) on transcripts from the VHL gene. The facets and colours indicate which time point in the HEK calibration dataset the data were taken from (from left to right: 0h, 4h, 6h, 8h, 12h). (B) Taking the data for VHL from A, the mean is shown as points and the 95% confidence intervals are shown as error bars. The data are coloured by time post-ActD as in A. (C) The number of fit sites per gene from the HEK Calibration experiment are shown as a histogram with a $\log_{10}$ transformed y-axis. (D) The relationship between the standard error of the mean gene ages and the number of sites per gene is shown. The black dotted line denotes the fit of a linear model to the $\log_{10}$-$\log_{10}$ transformed data. The colouring denotes the 2D kernel density. (E) The cumulative percentage of genes that have a 95% confidence interval of width x hours is shown for four different cut-offs for the minimum number of fit sites per gene. The vertical black dotted line indicates a 95% CI of width 1 hour. The x-axis is truncated for visualisation purposes: the dark blue line extends to a maximum value of 40.33 hours.

The performance of the per-site age method was assessed by applying it to the various calibration datasets as detailed in the previous Chapter (Chapter 2).

Since the mean gene age is calculated by taking the mean of the per-site ages, the accuracy should increase with the number of sites per gene (invoking the law of large numbers). Using the HEK calibration dataset (endogenous ADAR1, Table 2.1), this was investigated by analysing how the mean age estimates, standard error of the

mean (SEM) and variance changed for genes with different numbers of fit sites (Figure 3.2C,D). This showed that the SEM but not the variance was negatively correlated with the number of fit sites per gene (Figure 3.2D), indicating that genes with more sites did produce more precise mean age estimates but that the variance of the age estimates was unrelated to the number of sites.

To investigate the effect of two other variables on the precision of the mean age estimates - the number of reads at the sites and the editing rates (λ_i) - a multiple log-log regression was performed of the form

$$\log(\text{SEM}) = \beta_0 + \beta_1 \log(\text{n\_sites}) + \beta_2 \log(\text{Reads}) + \beta_3 \log(\text{Lambda}) + \epsilon. \quad (3.3)$$

The fit model yielded an adjusted R^2 of 0.31, indicating that approximately a third of the variance in the SEM can be explained by these predictors (all had p values less than 10^{-8}). However, since the R^2 value was only moderate, the majority of the variance is explained by other factors.

Of the three coefficients of the model, β_1 was the largest in absolute magnitude (-0.31) suggesting that the number of sites is the primary determinant of the variability in the SEM. This was lower than the expected value $\beta_1 = -0.5$, which would be expected if the per-site age estimates, $\bar{\tau}_i$, were truly independent and identically distributed (invoking central limit theorem), indicating that one or both of these assumptions was violated. This may be explained by editing sites having different editing rates λ_i and thus different variances of their age estimates - meaning that they are not identically distributed. It may also be the case that editing at one site influences editing at another, yet this was not determined.

For the 1,940 genes that have at least two fit sites in the 0h time point (i.e. the cells at rest) from the endogenous HEK calibration experiment, the average SEM was $\pm$ 0.76 hours and 483 (24.9%) of these genes had a 95% confidence interval (95%

CI) for their mean age of 1 hour or less (Figure 3.2E). As shown by plotting the cumulative distribution of the number of genes vs the 95% CI in Figure 3.2E, if the threshold for the number of sites per gene was raised, then the proportion of genes with a 95% CI under 1 hour increased, as expected from the previous analysis of the SEM.

Having characterised the error of the mean gene estimates, these values were then compared across the four human datasets from Chapter 2: cortex, midbrain, HEK, and NLambda HEK. Since different editing sites were detected across the datasets (Figure 3.3A), it was first determined whether to subset the analysis to only shared sites (i.e. sites shared in all four datasets ($n = 8,096$ fit sites)) or to include all fit sites within each dataset. It was assumed that using only shared sites would produce age estimates that correlated more highly between the datasets, whilst decreasing the number of genes that were measured.

Investigating the overlap of the fit sites between the different calibration datasets, found a large difference in the counts of unique sites for each dataset (% of unique sites: cortex = 74.9%, midbrain = 21.0%, HEK = 28.6%, NLambda = 33.2%) with the cortex dataset being an outlier with a particularly high count of unique sites. This was likely due to this dataset being sequenced with the greatest number of reads.

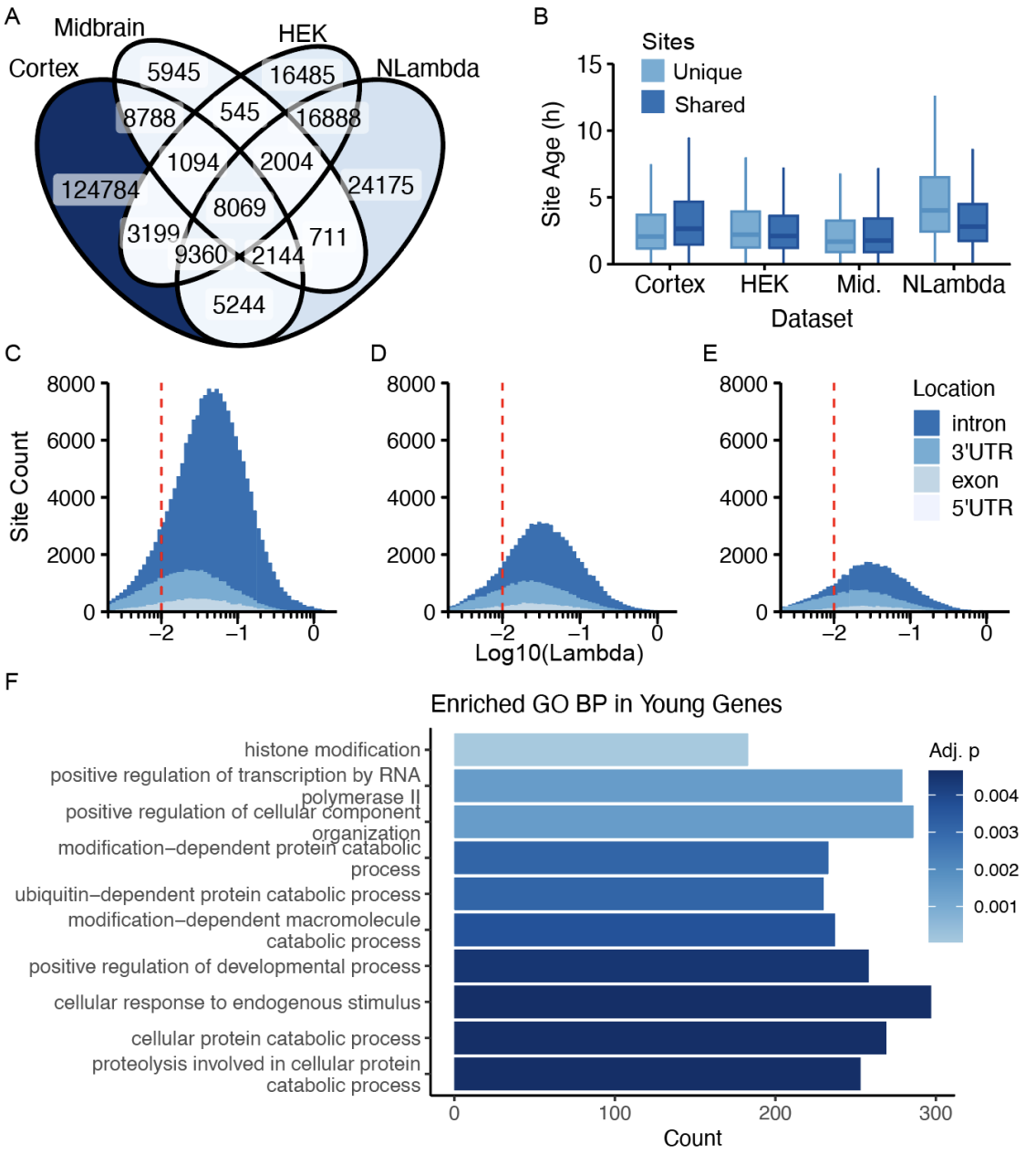

Figure 3.3: **The detection of new editing sites is sensitive to sequencing depth.** (A) Venn diagram showing the shared and unique sites between the four major calibration datasets as detailed in Table 2.1. (B) The per-site ages calculated from the 0h time point from each of the calibrations are shown as box-plots. The light blue box-plots show sites that are only found in the dataset specified on the x-axis whilst the dark blue shows sites that are found in at least one other dataset. The editing rates from the cortex dataset (C) were re-fit after downsampling the .bam files (D) to the same average depth as the midbrain dataset (E). The red line indicates the minimum editing rate threshold for fit sites. (F) Bar plot showing the top enriched gene ontology (biological process) (GO BP) terms among young genes (defined as having an age less than the median gene age), highlighting dynamic processes such as transcription control, protein modification and response to stimuli. Bar lengths show the number of genes annotated with each GO term. Significance was determined using a hypergeometric test with Benjamini-Hochberg correction for multiple testing ($p < 0.05$, $q < 0.2$) and the background universe of genes was restricted to genes containing fit sites.

It was reasoned that sites that are shared between datasets are likely to be highly expressed and in more universally expressed genes. Due to the filtering criteria discarding sites that had an editing rate less than 0.01h^{-1}, some drop-off around this threshold that would result in non-shared sites having lower editing rates was also expected. For each dataset in turn, I tested whether the shared sites and non-shared sites had significantly different read counts and editing rates (significance determined by Wilcoxon rank sum test). Across all four datasets, shared sites had significantly more reads than sites only found in a single dataset and for three of the four datasets, shared sites also had higher editing rates (the exception being the midbrain dataset which had no significant difference ($p = 0.21$)) (Figure 3.3B). This suggested that the sequencing depth of the samples had a non-trivial effect on the detection of sites and thus on downstream age prediction with the filtering criteria that was chosen. To demonstrate this, the cortex dataset was downsampled to have the same number of reads as the mean of the midbrain dataset and the two compared (Figure 3.3C,D,E). The downsampled dataset produced 63,148 fit sites - a 61.2% decrease from the original cortex dataset ($n = 162,741$ fit sites) - but which was still more than double the midbrain dataset ($n = 30,154$). Proportionally, the intronic fraction exhibited the steepest decrease, dropping from 82.2% of fit sites to 67.8%. All other fractions increased in proportion.

Finally, it was investigated whether there were biologically meaningful differences between the genes that had old mean ages compared with those that had young mean ages. Taking the 0h time point from the cortex calibration experiment, I calculated the mean ages for all genes present (n = 6,977) and split the dataset in half about the median gene age value (2.99 hours), labelling genes with a mean age less than median as 'young' and genes greater than median as 'old'. Using `clusterProfiler`, the young and old sets of genes were compared by their Gene Ontology Biological Process (GO BP) annotations which identified a number of significantly enriched processes in each group. The results showed that younger genes were enriched for pathways related to regulatory and epigenetic functions, such

as histone modification and RNA splicing, whereas older genes were associated with house-keeping processes like ribosome biogenesis and several metabolic pathways. Repeating the analysis with the Reactome database gave similar results, with a strong enrichment for cell cycle pathways in young genes. Both of these results were in agreement with the literature that suggests that genes with faster turnover are more likely to be involved in dynamic processes such as transcription regulation and signal transduction whereas older genes are likely to be less dynamic and more involved in core cellular functions such as metabolism (Yang et al., 2003). However, restricting the background set of genes used by `clusterProfiler` to the set of genes with mean age estimate (rather than the whole genome) resulted in no significant results for old genes, although the enrichment of regulatory pathways in young genes persisted (Figure 3.3F). These results suggested that whilst there are pathways that young genes are significantly enriched for, old genes are not as strongly clustered in specific processes and pathways.

In conclusion, the above results demonstrated that estimating mean gene ages from individual sites produced accurate estimates and that the precision of these estimates increased with the number of sites per gene. Having established this, it was then asked whether it would be possible to determine not just the mean age of transcripts, but the ages of individual transcripts.

3.3 Per-transcript Age

The theoretical framework for estimating the ages of single-molecules (i.e. transcripts) was previously introduced in Chapter 2 Section 2.2. Conceptually, as compared with per-site age - which calculates the age at a *single* site *over* a set of transcripts - per-transcript age estimates the age of a *single* transcript *over* its sites. Unlike per-site age, which is constrained to estimating the first mean of the distribution of transcript ages, per-transcript age in principle can estimate the full distribution. An approximation of the full distribution of transcript ages for a given

gene would yield richer information on past transcriptional changes such as increased or decreased transcription in the recent past.

Per-transcript age depends on being able to phase editing sites from RNA sequencing data: that is, to know that a set of base calls in the data resulted from the same cDNA fragment from the library preparation. If editing sites are very close together, then this would be possible with short-read sequencing, as has used in the work reported so far. However, in practice, endogenous editing sites are too dispersed in sequence space for more than several sites to be phased by short-read sequencing (the mean distance between fit sites in the 0h time point from the NLambda HEK calibration dataset is 200.5 base pairs (bp), with a standard deviation of 382.3 bp). We therefore used the long-read sequencing platforms offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT).

Although a single transcript may produce several reads due to polymerase chain reaction (PCR) amplification steps when preparing the cDNA libraries for sequencing, the terms 'transcript' and 'read' are used interchangeably in reference to age estimation (as explained at the end of Chapter 1 Section 1.8.1). However, under the assumption that all transcripts for a particular gene are amplified proportionally, the total number of reads produced will not skew the distribution of transcript ages since the distribution is normalised. This invariance of the predicted ages to the number of sequencing reads mapping to a particular gene is a useful and non-trivial property, as it frees comparison of mean ages between samples from the RNA composition effect that plagues abundance comparisons (such as differential expression).

As introduced in Chapter 2 Section 2.2 , the likelihood of a transcript, r, being of age τ_r given the observed editing states at its sites, $\boldsymbol{x}$ is given by

$$L(\tau_r|\boldsymbol{x}) = \prod_{i\in G}(1 - e^{-\lambda_i\tau}) \prod_{i\in A} e^{-\lambda_i\tau}. \tag{3.4}$$

A maximum likelihood estimate (MLE) of the transcript age, $\hat{\tau}_r$, can be approximated numerically (as described in Methods Section 6.3.2).

In order to build a likelihood model for each read from long-read sequencing methods, two primary technical challenges had to be overcome. First, how to efficiently retrieve the individual base calls at editing sites for each read in turn and, second, how to handle the potentially very large number of likelihood models that could arise from any given dataset.

3.3.1 Extraction of phased edits from BAM files and single-molecule age estimation

During the course of the research described in this thesis, several extremely deep long-read datasets were generated, with one in Chapter 5 containing 1.2 billion reads across 18 samples (i.e. one PromethION flow cell per sample). The resulting BAM files were up to 99Gb in size and no publicly available software was found that would be able to extract base calls at specific sites from individual reads at this scale. Therefore, a new algorithm was developed to achieve this, which is implemented in R4.1.1 (Algorithm 1) and detailed below:

Algorithm 1 RNA Editing Call Extraction Pipeline

Input: BAM file, IsoQuant read-to-gene mapping, editing rates
Output: The base calls at editing sites on a per-read basis
Define `chunk_size` as 10^6 reads
for each BAM file in `BAMS` **do**
    Create output directory if not exists
    Load IsoQuant read-to-gene mapping
    Filter reads to those mapping uniquely to genes containing fit sites
    Open BAM file with `chunk_size` yield
    Initialise `counter` to 1
    while BAM file has more reads to process **do**
        Read next chunk of BAM data
        Extract editing calls using `cigar_decoder_ranges_multi`:
            Parse CIGAR strings to determine read alignment structure
            Convert read alignments into genomic coordinate ranges
            Identify reads overlapping with specified editing sites
            Extract corresponding base calls from read sequences
            Extract corresponding quality-scores
            Store base calls and quality scores as strings in a data.frame object
        Save extracted calls to temporary file
        Increment `counter`
    end while
    Close BAM file
end for
Read all temporary files back into memory
Concatenate all extracted call data into a single dataset
if filtering on quality-score or editing rate is specified **then**
    Apply filtering criteria to concatenated dataset
end if
Save final processed dataset to output file

This algorithm avoids exhausting memory by processing BAM files in chunks, maps reads to genes (using the read-to-gene mapping generated by IsoQuant (Prjibelski et al., 2022)), extracts the base calls at editing sites, and filters the results based on quality (Q-score) or an editing rate (λ_i) threshold. The core function `cigar_decoder_ranges_multi` optimises performance by using the `GenomicRanges` package to find multiple sites on multiple reads at once.

The resulting files are small enough to be stored in a human readable format and thus are stored as tab separated value (.tsv) files. An example is shown in Table 3.1.

Gene ID	qname	Calls	Site UID	QVs
ENSG00000164466	m64045...	AAAAA...	5_175527974_1/...	93/93/...
ENSG00000130203	m64045...	GAGAG...	7_98237482_1/...	90/92/...

Table 3.1: Example output from Algorithm 1. Column values are truncated for readability. 'qname' contains the read ID from the sequencing run, 'Calls' is a string of base calls at the editing sites found on the read, 'Site UID' is a forward-slash-separated string containing the unique identifiers of the site loci (Site UIDs) found on the read and 'QV' is a string containing the phred quality scores from the run.

Using the extracted editing profile of each read from the output of Algorithm 1, a likelihood function is constructed using Equation 3.4. In practice, this is done using a function factory, which produces a list of function calls with each function's enclosing environment containing the editing rates for the sites that are found on the read. This formulation is quick to generate and allows for downstream manipulation of the likelihood functions - such as normalisation - although typically only the MLE of the age of each read is required.

3.4 Determining transcript ages in the calibration experiments

Long-read sequencing was tested on four time points from the NLambda HEK calibration experiment (0h, 2h, 8h and 16h post-ActD addition) using the PacBio Sequel II system which was selected due to its high fidelity. The workflow described above (Algorithm 1 and the function factory) was used to obtain the MLE of the age of for each transcript. The data are summarised in Table 3.2

Time Point	Raw reads	Aligned Reads	Reads in edited genes	Reads with sites	Reads with >=5 3'UTR sites
0h	2,249,394	2,207,845	1,365,388	255,917	9,541
4h	2,632,694	2,581,052	1,617,945	340,154	13,740
8h	1,437,618	1,398,220	849,942	173,948	11,072
16h	2,276,961	2,228,454	1,389,423	293,129	18,759

Table 3.2: **Counts of long-reads from PacBio sequencing of the the NLambda HEK calibration experiment**. 'Raw reads' denotes the count of Hi-Fi reads in the fastq files. 'Aligned reads' denotes the primary alignments against GRCh38.100 using minimap2. 'Reads in edited genes' denote the count of reads annotated to well-edited reference genes by IsoQuant (Prjibelski et al., 2022). 'Reads with sites' denotes the counts of reads with one or more editing sites detected. 'Reads with $\geq$ 5 3'UTR sites' denotes the number of reads annotated to genes that contain at least five fit sites in their 3'UTR.

As shown in the Table 3.2, only a small fraction of the total reads in the dataset contained fit sites and fewer still had multiple sites in the 3'UTR. This motivated the use of Oxford Nanopore Technologies (ONT) sequencing, which offers higher-throughput, over PacBio in future experiments.

Similarly to how the mean gene age, $\bar{\tau}_g$, for the per-site age estimates was calculated by taking the mean over the sites (Equation 2.11), an estimate of gene age was also calculated from the per-transcript age estimates by taking the mean of the MLE values, $\hat{\tau}_r$ of all the reads assigned to a given gene (Methods Section 6.3.2). Since there was significant variability in the number of reads and editing sites across the genes in the dataset, the variance and the error in the MLE estimates was investigated.

3.4.1 Many editing sites discovered during calibration are absent from long-reads

For the NLambda HEK calibration data, 4,644 genes (out of the 6,386 genes with fit sites from the short-read sequencing data) were also present in the PacBio long-read sequencing data. Across these genes, the mean number of reads with at least one editing site was 55.1 reads per gene, with a standard deviation of 369.9. 60.1% of these reads had a single editing site, whilst 9.63% had 5 sites or more, 2.5% had 10 or more and 0.89% had 20 or more: a sharp drop-off. Obtaining only 24,634 reads with five or more editing sites across 1,230 genes from a sequencing run of 2,249,394 reads (a 1.10% retention) was surprisingly poor, especially since the mean number of fit sites identified from the calibration experiment for those same 1,230 genes was 54.6 sites per gene - an order of magnitude higher.

To investigate potential causes for the absence of sites from long-read sequencing data, long-read and short-read sequencing data of the same sample (the 0h time point from the NLambda HEK calibration) were compared. In the long-read data, reads are assigned to genes using IsoQuant (Prjibelski et al., 2022) and so for each

gene there is a set of reads that are expected to contain the fit sites. For each fit site, I calculate the frequency with which it appears on those reads. On average, sites were only present on 22.6% of the reads that would be expected to contain them.

Splitting the sites into groups based on the regions of the gene they are annotated to showed that intronic sites were most commonly missing, being observed on an average on just 12.0% of reads. For the other regions, sites had mean frequencies of 28.2% for 5'UTR sites, 33.7% for 3'UTR sites and 40.7% for exonic sites. Additionally, 66.4% of fit sites were not detected on any long-reads at all, which may be due to the far lower sequencing depth of the PacBio sequencing compared with the short-read Illumina sequencing that was used when performing the calibration experiment. In support of this hypothesis, sites that were totally absent in the long-read data tended to be on more lowly expressed genes than sites that were detected (mean number of reads per gene = 6.84 vs 19.08 for absent vs detected sites. Corresponding standard deviations were 21.95 and 118.33 reads). However, visual inspection of the sequencing data and comparison with the Illumina sequencing of the same samples suggested that there may be other issues with either the alignment step (performed by minimap2) or with the assignment of reads to genes (performed by IsoQuant). A full analysis of these differences was beyond the scope of this work.

With these results demonstrating that a large proportion of PacBio long-reads contained very few editing sites, how this drop off affected the estimates of transcript age was investigated. In principle, the more sites a read has, the less sensitive the MLE, $\hat{\tau}_g$ should be to the editing state of any single site (Figure 3.4A). This is similar to the decrease in error of the mean gene age with increasing number of sites found for the per-site ages (Figure 3.2D).

Estimating transcript ages from the PacBio sequencing of the NLambda HEK calibration time points revealed a high frequency of reads for which $\hat{\tau}_g = 0$h, which are produced when all of the sites on a read are unedited - these reads are termed 'zero-edits'. For the unfiltered long-read data (minimum sites per gene = 1, minimum

reads per gene = 1), the data was dominated by zero-edits, with 91.8%, 87.8%, 87.5% and 86.0% of the reads 0h, 4h, 8h and 16h post-ActD being zero-edits, respectively (Figure 3.4B).

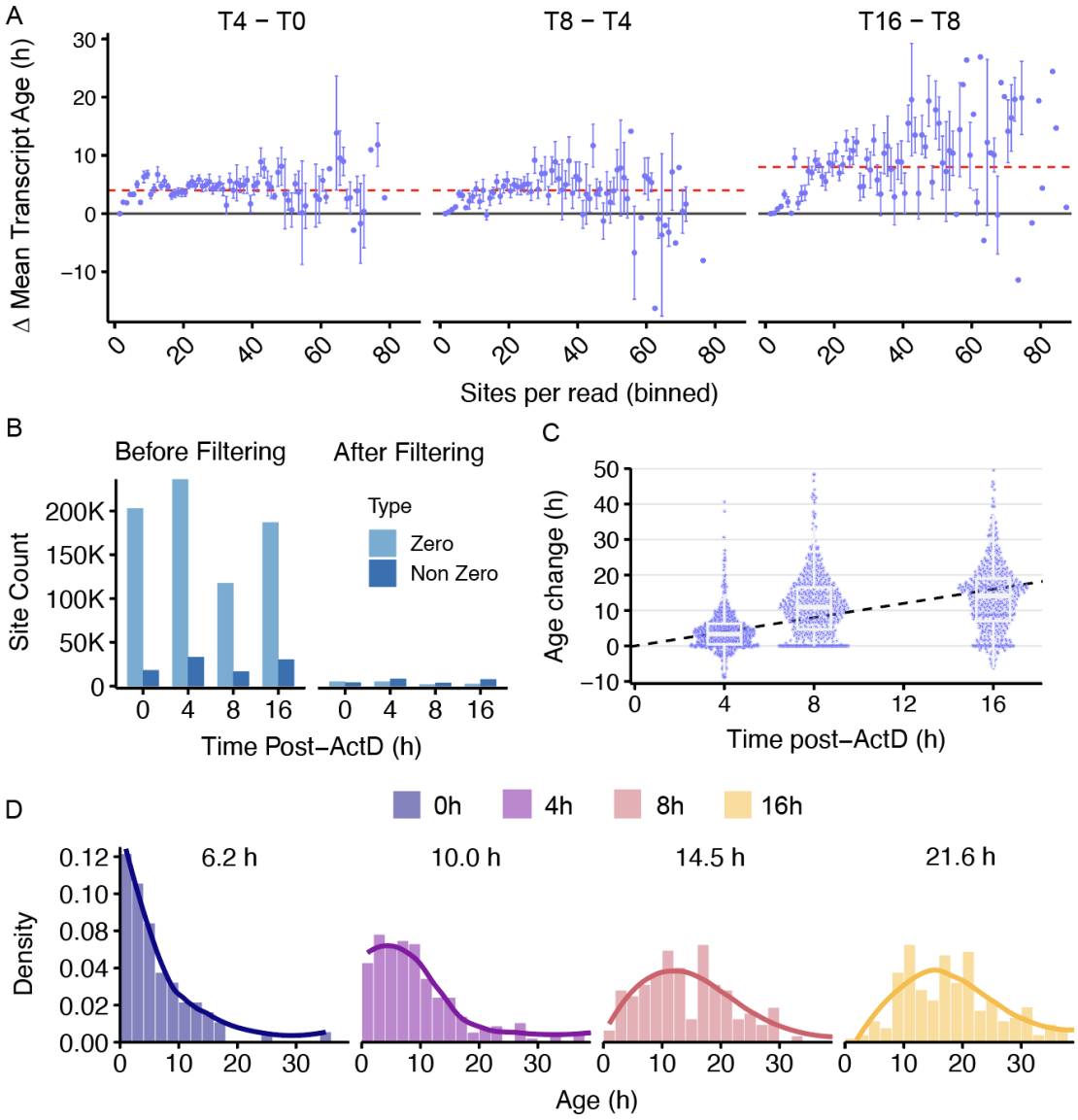

Figure 3.4: **Estimating the age of single transcripts from PacBio long-reads** (A) The maximum likelihood estimates (MLE) of transcript age are split into bins by the number of sites per read and the difference in age between time points (specified at the top of each facet) is shown. Points show the difference between the mean of the estimated ages of each bin between the two conditions and error bars show the propagated standard error of the mean. The red dashed line indicates the expected ΔAge. (B) The counts of unedited reads (zero-edits) and non-zero edits are shown before and after filtering sites to 3'UTR sites in genes with at least five sites. (C) Pairwise differences in the estimated gene ages are shown for 4h, 8h and 16h versus the 0h time point. The black dashed line is $y = x$. Box-plots shown 1st, 2nd and 3rd quartiles. Whiskers extend to 1.5 times the IQR in either direction. (D) The distributions of individual GATC transcript ages are shown as histograms for each time point. The estimated gene age is given at the top of each facet and the histograms coloured by the calibration time point from which the estimates were generated, showing the effect of ActD.

Under the model of mRNA birth and decay presented in Chapter 1 Section 1.3, the transcripts from a gene at steady state expression are expected to follow an exponential decay distribution with a modal age of 0 hours. Yet even with this assumption, most genes sequenced had a far higher number of zero-edit reads than expected. Even 8 and 16 hours post-ActD addition, there was still a substantial number of zero-edit reads suggesting that these reads are not amenable to age estimation at all (Figure 3.4B,C). To investigate if there was a relationship between the number of sites per read and the presence of persistent zero-edit reads, I split the set of reads into bins by the number of sites present on each read (in increments of one site) and plotted the mean estimated transcript age for each bin. Inspecting the differences (ΔMean Transcript Age) between the 0h, 4h, 8h and 16h post-ActD revealed that the mean estimated age for reads with a single site was 0 hours for every comparison and that the difference in the mean estimated age between time points was smaller than expected for reads up to 5 sites (Figure 3.4A). The implication of this is that reads with few editing sites are poor candidates for age estimation, since they show little change in age over the several hours between sequencing steps.

Time-invariant reads were filtered from the dataset by subsetting the list of fit sites to only 3'UTR sites in genes with at least five 3'UTR sites. This heuristic was used to select sites with a well-known basis for editing - being present in *Alu* repeats (introduced in Chapter 1 Section 1.4). This substantially reduced the calibration dataset from 201,233 sites ($n = 8,685$ genes) to 32,803 sites ($n = 949$ genes) (Figure 3.4B) but improved performance. Using this list of fit sites, the percentage of zero-edit decreased dramatically for each of the samples from 91.8% to 55.4%, 87.8% to 38.8%, 87.5% to 33.4% and 86.0% to 24.9% of the reads 0h, 4h, 8h and 16h post-ActD being zero-edits, respectively (Figure 3.4B). Although a large fraction 'zero-edits' was persisted in the 8h and 16h time points - many hours after halting transcription.

The accuracy of the per-transcript age method was measured by linear regression, similar to the per-site method in Chapter 2 Section 2.3 (Figure 3.4C). For each

gene g, an estimate of the gene age $\overline{\hat{\tau}}_g$ was calculated from the transcript MLEs, $\hat{\tau}_r$ for each of the calibration time points. The estimated ages were expressed as age changes, $\Delta\overline{\hat{\tau}}_g$, relative to the age in the 0h condition. A linear model was fit to each gene in turn, yielding an average gradient of 0.78 with a mean R^2 of 0.53 over the genes tested ($n = 601$) (Figure 3.4C). This was a worse performance than the per-site age estimates obtained from the Illumina sequencing of the same samples which had a gradient of 0.90 and mean R^2 of 0.83. Further analysis showed that of the 601 genes that were fit with the linear model, 69% of them had reads in all four calibration time points, 12.8% were found in three time points and the remaining 18.1% were found in just two time points. This indicated again that sequencing depth - and subsequent drop out of genes with few reads - was a contributing factor to the poorer performance.

In conclusion, these results demonstrate that the age of individual transcripts can be estimated from PacBio long-read sequencing and that the distributions of the transcript ages reflect transcription shut-off (Figure 3.4D).

3.4.2 ONT long-reads are high accuracy and can be used for transcript age estimation

The sequencing depth offered by PacBio severely limited the analysis in the previous section. Therefore, Nanopore sequencing (ONT) was used for future experiments. Due to the lower reported accuracy of ONT sequencing, a rigorous characterisation of the per-base Q-scores was performed to ensure that the platform would not introduce excessive error rates.

ONT offers several sequencing products, which can most simply be grouped by the flow cell used for sequencing (as introduced in Chapter 1 Section 1.6.2). The three groups, in increasing order of yield, are: Flongle, MinION and PromethION. In 2022, ONT released the R10.4.1 chemistry which provided a substantial improvement over the previous R9.4.1 - with R10.4.1 reporting a median Q-score per-read of at least

Q20 (corresponding to a 99% base calling accuracy).

The performance of ONT sequencing was tested on the 1h and 9h time points from the NLambda HEK calibration experiment, which were sequenced on a MinION and a PromethION24 using R10.4.1 flow cells[1].

The two samples were run on a single MinION R10.4.1 flow cell, generating 3.5 million reads and 9,579 transcript MLEs (Figure 3.5A). Inspecting the mean per-read Q-scores of the sequencing data with FastQC (Andrews, 2010) and NanoPlot (De Coster and Rademakers, 2023) surprisingly produced different results. The cause of this was determined to be that the per-read mean Q-score calculated by FastQC does not account for the logarithmic nature of phred quality scores. Instead of transforming the phred scores into error probabilities, taking the arithmetic mean and then converting back to phred scores, FastQC simply takes the mean of the phred scores. This resulted in higher reported Q-scores than were actually present. To avoid any other erroneous calculations, I calculated the mean, median and mode of the Q-scores of the base calls at the individual editing sites with custom code, rather than reporting per-read statistics. For the base calls at editing sites, this produced a mean Q-score of 18.92, median of 31, mode of 38 and a standard deviation of 10.13 (Figure 3.5B). By way of comparison, calculating the mean and median directly on the Q-scores without converting back to error probabilities (as FastQC does) yielded a mean of 29.65 and a median of 31, highlighting the issue with the FastQC mean and the robustness of the median.

[1]The 1h and 9h post-ActD conditions were sequenced due to the cDNA library for the 0h, 4h, 8h and 16h time points having been depleted from the previous Illumina and PacBio sequencing runs.

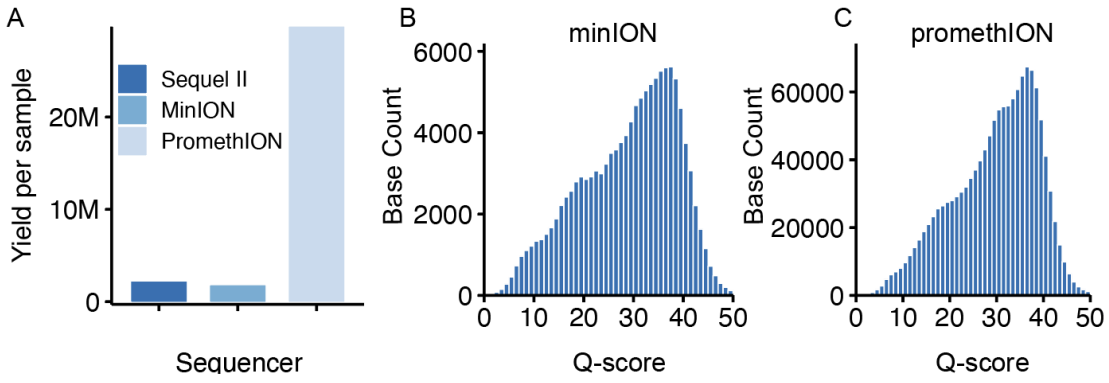

Figure 3.5: **Yield and accuracy comparison between the Sequel II, MinION and PromethION sequencing systems**. (A) Total number of reads produced by each sequencing platform: Sequel II (PacBio), MinION (ONT), and PromethION (ONT). PromethION produced substantially higher yield compared to the other platforms. (B-C) Distribution of base quality scores extracted at fit sites (Q-scores) for MinION (B) and PromethION (C) reads. PromethION sequencing had by the far highest yield per sample and a high average Q-score at bases of interest (fit sites), skewed toward higher-quality reads compared to MinION.

The analogous mean, median and mode for the Q-scores from the PacBio sequencing described above were 80.84, 88.62 and 89.44 - which are extraordinarily high due to the circular consensus sequencing method employed by PacBio. In practice, the median Q-score of 31 from the MinION was sufficient for editing analysis and the true error rates in the PacBio data are probably under-reported due to mutations from PCR (as recently shown by Sun et al. (2024)).

The same samples were run on a single PromethION flow cell, generating 59.54 million reads and 909,131 transcript MLEs per sample, a substantial increase over both the MinION and the PacBio output. The Q-scores at the editing sites showed a slight improvement over the MinION, with a mean Q-score of 20.39, median of 32, mode of 37 and a standard deviation of 9.40 (Figure 3.5C).

Repeating the analysis of the PacBio data for the PromethION data (which had a 15-fold higher depth) increased the number of detected genes from 4,644 genes to 7,331. The PromethION data had a mean of 174.9 reads per gene (SD = 2,333.9) compared with 55.1 reads per gene (SD = 369.9) for PacBio. The occurrence of zero-edit reads was largely unchanged, with 91.4% for the 1h post-ActD condition

and 86.4% for the 9h. Performing the same filtering as for the PacBio data (only 3'UTR sites are retained and reads must contain at least five sites) dropped the rate of zero-edits drops to 38.3% and 18.6% for the 1h and 9h respectively - a larger decrease than observed for the PacBio data.

Using the PromethION data, the ages of almost a million transcripts could be estimated in each sample, with a mean estimated gene age of 4.9 hours in the 1h post-ActD sample and 14.0 hours in the 9h post-ActD sample ($n = 733$ genes) (Figure 3.6A). For each gene, the difference in age estimate between the 9h and the 1h, $\Delta\overline{\hat{\tau}}_g$, was calculated which had a mean of 9.00 hours over all the genes (SD = 8.72, median = 8.81 hours and IQR = 6.11). Both the mean and the median were larger than the known difference in time between the samples (8 hours), which suggests a bias towards overestimating the ages of old genes (Figure 3.6B).

A Q-Q plot of the data showed that most of the data was distributed normally but deviated substantially at the outer quantiles - suggesting that some genes performed poorly. Poor performance had a moderate inverse correlation with the number of transcripts, suggesting again that read depth may be limiting (Pearson's correlation coefficient was -0.35). Another metric, termed 'mean states' was calculated which was defined as the mean number of unique MLE values per condition per gene (Figure 3.6C). Intuitively, this metric integrates both the number of transcripts and the number of sites per transcript. The equivalent Pearson's correlation coefficient for $\log_{10}$(mean nStates) was -0.38, making it a slightly better predictor of $\Delta\overline{\hat{\tau}}_g$. Of note, all of the genes that had $\Delta\overline{\hat{\tau}}_g = 0$h (i.e. - completely time invariant) had only one age state in both conditions but some had multiple transcripts - suggesting that filtering the data by the number of states rather than the number of transcripts may be more effective.

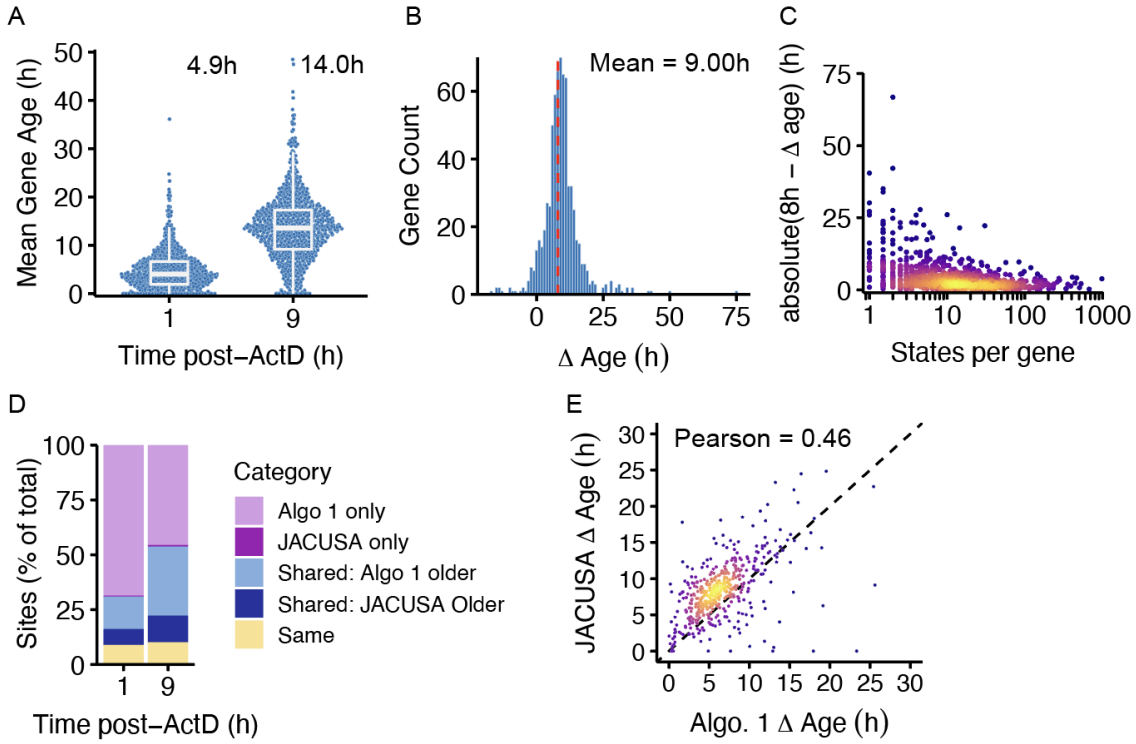

Figure 3.6: **Transcript ages encode known changes in time**. (A) Gene ages are estimated from PromethION long-read data by taking the mean of the maximum likelihood estimates (MLEs) of age of individual transcripts. The mean of the estimated gene ages for the 1h and 9h post-ActD samples are shown at the top of the plots. Box-plots show the 1st, 2nd and 3rd quartiles with whiskers extending to 1.5 times the IQR. (B) The ΔAge for each gene is calculated by subtracting the estimated age in the 1h sample from the 9h sample. The dashed red line shows ΔAge = 8h - the expected value: the observed average is 9.00 hours. (C) The relationship between the number of unique MLE values for a gene and its absolute deviation from the expected ΔAge shows that genes with more states more accurately display the expected age difference. (D-E) The difference the editing data retrieved by JACUSA2 and Algorithm 1 for site ages (C) and the ΔAges obtained from transcript age estimates (D).

3.4.3 The editing information retrieved by JACUSA2 differs from that retrieved by Algorithm 1

Whilst long-read sequencing is required for per-transcript age estimation, the converse - that short-read sequencing is required for per-site age estimation - is not the case: per-site age estimates can be calculated from long-read sequencing data. Since short-read sequencing is far more commonly used than long-read, showing that the age estimates agree between the two was of interest. Therefore, an analysis of how

well the per-site and per-transcript age methods correlated on the same long-read data, and how the per-site method correlated *between* sequencing modalities was performed. In total, four comparisons were made:

1. Per-site (long-read, JACUSA2) vs per-site (long-read, Algorithm 1)
2. Per-site (long-read) vs per-transcript (long-read)
3. Per-site (short-read) vs per-transcript (long-read)
4. Per-site (short-read) vs per-site (long-read)

Comparison 1: Per-site (long-read, JACUSA2) vs per-site (long-read, Algorithm 1).

The per-site and per-transcript methods use different pre-processing pipelines to retrieve editing information from sequencing data. Per-site uses the editing levels produced by JACUSA2 (Piechotta et al., 2021) whereas the per-transcript uses Algorithm 1, from which editing levels can be manually calculated. The extent to which the choice of preprocessing pipeline affected the results was investigated.

On the same PromethION data as used above, JACUSA2 produced 16,181 fit sites across the 1 h and 9 h samples, whereas Algorithm 1 produced 29,362 sites (Figure 3.6D). 15,994 sites were shared between the two datasets, of which 71.4% in the 1h time point had identical values for both the number of edits and the number of reads at the shared sites. This increased to 81.2% of the sites in the 9h condition. The lower number of sites detected by JACUSA2 - especially in the 1h condition - was largely driven by the minimum depth threshold (five reads at any given site) and minimum editing threshold (one edit) implemented by the software. Applying the same thresholds to the Algorithm 1 output reduced the number of sites from 29,362 to 18,591 for the 1h condition, which was closer to the number observed from JACUSA2.

The accuracy of the age estimates was judged (as before) by how closely they predicted the known 8 hour gap between the 1h and 9h sequencing steps - referred to

as the ΔAge. ΔAge was calculated at 7,446 sites identified by both JACUSA2 and Algorithm 1. Only 7.42% of these sites had the exact same ΔAge value from both methods and the ΔAges calculated from JACUSA2 showed a bias towards being larger. Calculating the mean of the site ages per gene (rather than just the individual sites) produced a smaller difference between the two methods, with the results given in Table 3.3 and Figure 3.6E.

Method	Genes	Mean ΔGene Age	SD ΔGene Age
JACUSA2	698	5.79	5.71
Algo 1	829	7.26	6.17

Table 3.3: **Differences in per-site gene age estimates between the 9h and 1h conditions for edits extracted by JACUSA2 and Algorithm 1**. The difference (Δ) in estimated gene age between the 9h and 1h condition is shown for the editing information extracted at sites either by JACUSA2 or by algorithm 1, with the ΔGene Age from Algorithm 1 being closer to the expected value of 8 hours. The data is not subset to shared sites between the two methods. SD indicates standard deviation.

Ultimately, since the difference in estimated gene ages calculated from the output of Algorithm 1 was closer to the true elapsed time between lysis of the two samples (8 hours) and JACUSA2's filtering criteria were an unnecessary source of bias; the editing information extracted from algorithm 1 was used going forwards[2]

3.4.4 Gene ages estimated from the per-site and per-transcript methods correlate well

Comparison 2: Per-site (long-read) vs per-transcript (long-read).

I next compared the age estimates from the per-site and per-transcript methods on the same PromethION sequencing data. The gene ages correlated well both in the 1 h ($n = 641$ genes) and 9 h ($n = 706$ genes) time points with Pearson correlation coefficients of 0.77 and 0.75 respectively (Figure 3.7A,B). The ΔAge for each of the

[2]This discovery has implications for the per-site estimates generated from short-read sequencing data, for which I only extract the editing information by JACUSA2. The likely effect is a bias towards estimating older ages, since unedited sites will be excluded.

methods correlated more weakly, with a Pearson correlation coefficient of 0.51 (n = 733 genes) (Figure 3.7C).

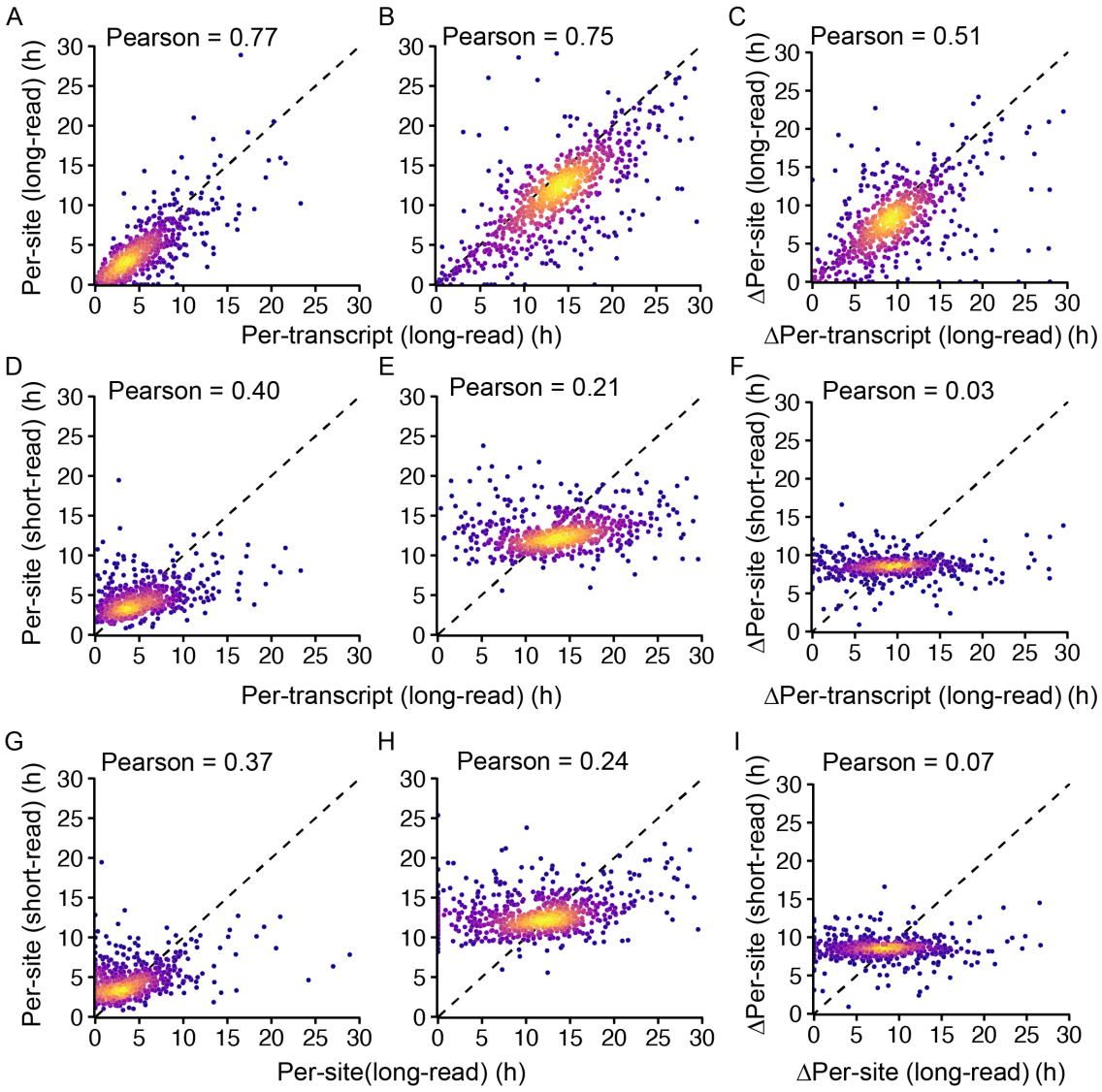

Figure 3.7: **Comparison of estimated gene ages derived from per-site age modelling and per-transcript age modelling on short-read and long-read sequencing platforms.** Correlation plots. Columns show the comparisons for: 1h (left), 9h (middle) and the 9h - 1h (i.e. the ΔAge) (right). Rows show the comparisons for: per-transcript vs per-site (both on long-read data, top), per-transcript (long-read) vs per-site (short-read) (middle) and the per-site (long-read vs short-read) (bottom). The colour shows the 2D kernel density of the plots, dashed black line shows $y = x$. Pearson values denote the Pearson correlation coefficient.

Comparison 3: Per-site (short-read) vs per-transcript (long-read).

The performance of the per-site age method between the Illumina and PromethION datasets was noticeably worse than when both methods were applied to the long-read

data (Figure 3.7D,E,F). Almost no correlation was observed for the ΔAge between the two methods on the two sequencing modalities (Pearson correlation coefficient = 0.03).

When taken in isolation, each of the methods appeared to accurately predict the ΔAge, with the mean from Illumina being 7.78 hours and the mean from ONT being 9.05 hours (compared to the groundtruth of 8 hours). However, the standard deviation of the PromethION data was much larger than the Illumina (SD = 8.55 vs SD =2.52) - which can be seen in Figure 3.7F) - and mostly likely produces the very poor correlation.

One possible explanation for this is that the editing information from the Illumina data was generated with JACUSA2 and - as discussed above - JACUSA2 will not report unedited sites or those with fewer than five reads. As can be seen in both Figure 3.7D and Figure 3.7G, there is a distinct lack of gene ages in the range 0 to 2 hours for Illumina gene age estimates, which may be in part due to under-representation of unedited sites that are missed by JACUSA2. It may be possible that adapting Algorithm 1 for Illumina BAM files would produce better agreement between the ΔAge values.

Comparison 4: Per-site (short-read) vs per-site (long-read).
Finally, the per-site gene ages from the PromethION data and the Illumina data were compared (Figure 3.7G,H,I), which produced similar results to Comparison 3.

In summary, these comparisons suggested that the choice of sequencing technology produced greater variation in model accuracy than whether the per-site or per-transcript method was used. In general, the dispersion of gene ages was greater for PromethION data than Illumina data but comparing the two age calculation methods showed that they were more consistent when applied to the same sequencing dataset. Even though the PromethION system offered far greater depth per sample than the Sequel II or the MinION, the depth used here (29.5M reads per sample) was still limiting. In the next section, a deeper PromethION dataset is produced

that improved performance.

3.5 Gene ages provide an estimate of mRNA half-life

The model for mRNA synthesis and degradation introduced in Chapter 1 Section 1.3 has a zero order synthesis rate parameter and a first-order degradation rate parameter. Under these conditions, the concentration of mRNA molecules decays exponentially over time, according to the degradation rate constant, δ. The mRNA half-life, $t_{1/2}$, is given by

$$t_{1/2} = \frac{\ln(2)}{\delta} \tag{3.5}$$

which follows from the definition of exponential decay.

At steady state, the rate of mRNA synthesis equals the rate of mRNA degradation and whilst the total number of mRNA molecules is constant over time, the individual transcripts are continuously replaced. In this system, the mean age, of the mRNA molecules is determined solely by the degradation rate, which is equal to $1/\delta$. Thus, if the mean age of the mRNA transcripts is known, the half-life is simply

$$t_{1/2} = \ln(2) \times (\text{mean age}). \tag{3.6}$$

As was introduced in Chapter 2 and has been discussed at length in this chapter, A-to-I editing encodes time and can be used to estimate mean ages of genes and therefore also mRNA half-life. In addition to the two methods introduced already - per-site mean gene age and per-transcript mean gene age - the half-life can also be calculated from the decay rate constant, δ obtained by fitting the distribution of MLEs with an exponential model of the form

$$N(\tau) = N_0 e^{-\delta\tau}, \tag{3.7}$$

where N is the count (or normalised count) of reads with MLEs of a certain age, τ is the age of the transcripts, δ is the degradation rate and N_0 is the count (or normalised count) of the youngest reads.

Half-lives were investigated with a new long-read dataset that my collaborator Ali generated. This dataset investigated the response of HEK293 cells (without the NLambda construct) to heat shock. Whereas the PromethION sequencing of the NLambda HEK calibration 1h and 9h samples (analysed in Sections 3.4 to 3.4.4) yielded bam files approximately 25GB in size, the new experiment produced bam files 55GB in size per individual replicate. Three biological replicates were sequenced per condition and two conditions were analysed here: the time point prior to heat shock (0h) and the control condition (no heat shock, cells were simply cultured at the resting temperature for a further 4 hours and then sequenced) - referred to as the 4h_control.

All six of these samples could be pooled together and considered as a single condition since there was little differential gene expression between the 0h and the 4h_control samples (as determined using DESeq2 (Love et al., 2014)). This pooling enabled 252GB of PromethION bam files to be used for a single condition - roughly 10 times more data than used in the analysis in Section 3.4. Using the editing rates of only the fit sites found in well-edited 3'UTRs (introduced in Section 3.4.1 and shown in Figure 3.4B), the ages of 600,569 individual transcripts were estimated from this dataset, resulting in exquisite resolution for many genes (Figure 3.8A,B,C). Furthermore, the correlation between the per-site and per-transcript age methods when applied to this dataset showed a very strong correlation (Pearson correlation coefficient = 0.92) (Figure 3.8D).

For the 792 genes detected, the half-lives calculated from the per-transcript gene age estimates with Equation 3.6 had a mean of 2.4 hours, median of 2.0 hours, standard deviation of 1.76 hours and IQR of 2.1 hours. For the same genes, the half-lives calculated from per-site method (from the same PromethION data) had equivalent

statistics of 2.04, 1.71, 1.52 and 1.80 hours (Figure 3.8D). These results showed an improved correlation between the per-site and per-transcript methods than the previous comparison (Figure 3.7A,B,C) and also displayed a skew to longer half-lives for the per-transcript method, especially for genes that have longer half-lives in the per-site method as well.

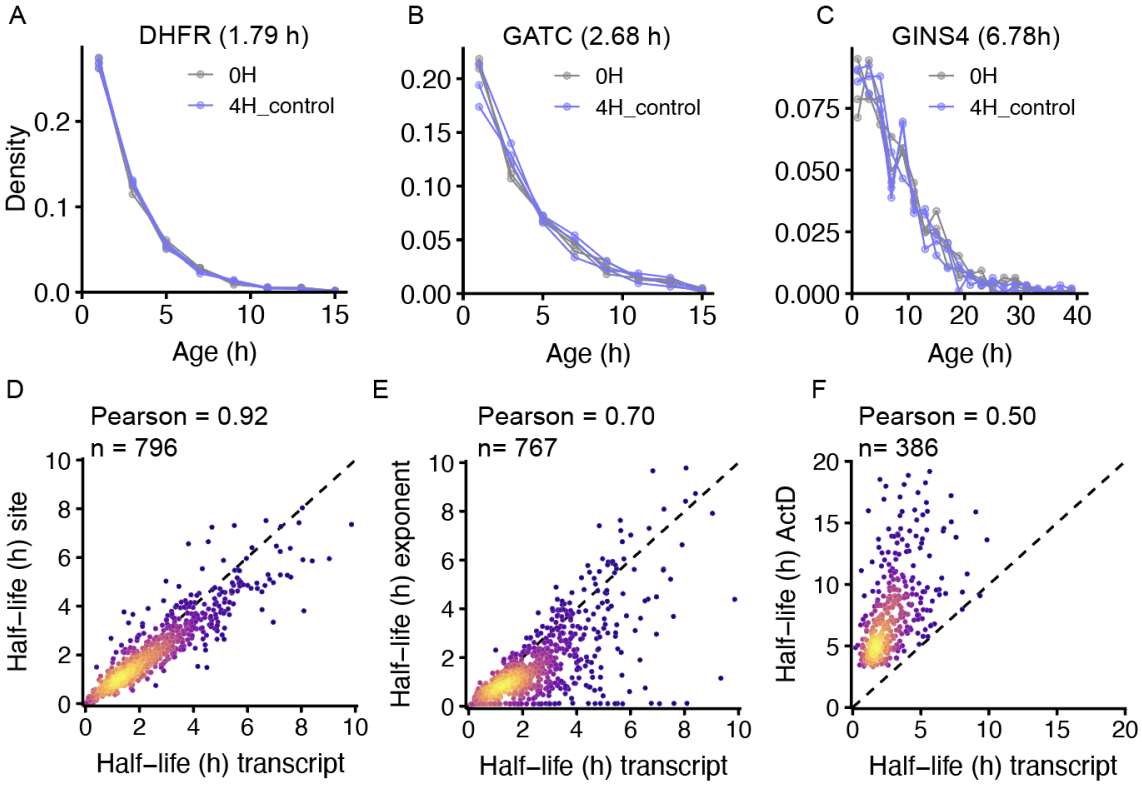

Figure 3.8: **Estimating mRNA half-lives from transcript ages in ultra-deep nanopore sequencing.** (A-C) Distributions of the estimated ages of transcripts for three genes of different half-lives, with each of the six samples plotted individually and coloured by the condition they were taken from in the HEK heat shock experiment (either the 0h or the 4h control (no heat shock) condition). (D - E) Correlation of half-lives from the per-transcript and per-site age methods (D), the per-transcript and the exponential fit to the MLE distributions (E) and the per-transcript vs the Viegas et al. (2023) protocol applied to the NLambda HEK calibration data.

Half-lives for 789 genes could be approximated by fitting Equation 3.7 to the distribution of transcript ages (such as those in Figure 3.8A-C), extracting the degradation rate parameter, δ and transforming into a half-life by equation 3.5 . The half-lives from this method were on average younger (mean = 1.72h, median = 1.30h) but exhibited similar dispersion (SD = 1.68h, IQR = 1.56h) to those from the per-site and

per-transcript mean ages. As shown in Figure 3.8D, the agreement of the per-site and per-transcript derived half-lives with those from fitting the exponential decay model (Figure 3.8E) supports the notion that the inferred ages of most transcripts follow the exponential behaviour predicted from a steady state model of mRNA turnover. However, a subset of genes observed in Figure 3.8E had a half-life of 0h predicted from fitting the exponential model. Inspection of these genes showed very few MLE values and a very sharp drop in the distribution of MLEs - which likely caused issues with the non-linear-least squares fitting method used.

Encouraged by the agreement between the three half-life calculation methods presented above, it was asked whether the half-lives from ERA correlated with common methods from the literature. The reported half-lives of human mRNA display significant variability between methodologies, cell lines and even laboratories making benchmarking challenging (as introduced in Chapter 1 Section 1.3.1). ERA half-life values were compared to several datasets from the literature (those used to train the ensemble model of Agarwal and Kelley (2022), namely the datasets from Tani et al. (2012); Schueler et al. (2014); Wang et al. (2014); Melé et al. (2017)), however, the partial coverage of the transcriptome of each dataset with the results presented here restricted multi-way comparison to a very small subset of genes.

Instead, the methodology of Viegas et al. (2023) - which estimates half-lives from time course RNA-seq after transcriptional shut-off from large scale normalisation of counts - was identified as a viable *in silico* method that could be adapted to the existing calibration datasets generated in Chapter 2. This method was adapted and applied to the NLambda HEK calibration dataset to enable benchmarking of the per-transcript and per-site half-life values against a method typical of the literature. Across all the genes tested ($n = 794$) there was a moderate correlation (Pearson correlation coefficient = 0.34) between the per-transcript method and the adapted method of Viegas et al. (2023). A subset of genes ($n = 386$), defined as having high temporal resolution ($\geq$ 70 states, $n = 386$ genes) showed an improved correlation between the two methods (Pearson's correlation coefficient = 0.50), although the

majority of half-life estimates derived from the per-transcript gene ages were younger than those of Viegas et al. (2023) (Figure 3.8F).

To investigate if there was a cause for the globally higher half-lives from the Viegas et al. (2023) method, the distribution of all the half-lives calculated with it ($n = 29,250$) showed a depletion of a half-lives around 0 hours. The consistently older half-lives produced by this method on our data may therefore be due to improper handling of low-count genes, especially those whose abundance (measured in transcripts per million) rapidly drops to 0 after just a few time points. Further refinement of this method on our datasets may improve the correlation between the age-derived half-lives and the Viegas et al. (2023) method.

In conclusion, the mean gene ages calculated from either the per-site or per-transcript ERA provide an estimate of the mRNA half-life. Whilst only applicable to well edited genes, the methods presented here have a distinct advantage of requiring neither chemical modification nor genetic engineering. Any future efforts that increase the number of genes with robust editing sites should also increase the number of genes for which the half-life of their mRNA can be calculated from these methods.

3.6 Discussion

The work of Rodriques et al. (2020) established that the ages of single transcripts could be inferred by counting the number of edits in a genetically engineered 3'UTR cassette. It was not known, however, whether the same principle would extend to the endogenous A-to-I editing in human cells of endogenously expressed and regulated human transcripts.

Here, I proposed a likelihood model (Equation 3.4) from which an estimate for the age of the transcript can be approximated by maximum likelihood estimation. The likelihood approach offers substantial advantages over the Poisson binomial used by Rodriques et al. (2020) (introduced in Chapter 1 Section 1.5) especially in the endogenous setting, where transcripts mapping to the same gene can have

different numbers of editing sites present (as shown in Section 3.4.1). Whereas the aim of Rodriques et al. (2020) was to estimate when a promoter driving the expression of a timestamped transgene was activated, that question is less applicable to endogenously regulated gene expression and instead the methods described in this Chapter can investigate properties of transcripts at or near steady state expression - as well as situations far from it (such as ActD-initiated transcriptional shut-off).

In addition to the per-transcript method, the novel per-site method for estimating the mean age of a gene was also introduced (Section 3.2. Unlike the per-transcript method, which requires long-read sequencing, the per-site method works on both short-read and long-read sequencing data. Given the current dominance of short-read sequencing, this feature makes the method far more accessible to future users and also enables retroactive application to publicly available datasets, although this was not demonstrated here.

The behaviour of the per-site and per-transcript methods was characterised here and identified shared and unique challenges. Both per-site and per-transcript methods are limited by the number of editing sites, which reduces the total number of genes that can be modelled as well as limiting the precision of the age estimates for the genes that do have editing sites. Both methods are also limited by the sequencing depth used, with this presenting a particular challenge for long-read sequencing which is more costly than short-read alternatives. Fortunately, the release of the high accuracy R10.4.1 sequencing platform from Oxford Nanopore Technologies during the course of this research provided a route to affordable ultra-high depth long-read sequencing which is necessary for accurate time resolution. The steep drop-off from the total number of reads used to sequence a sample and the number of transcripts for which ages can be estimated makes the method inherently wasteful of data. Whilst it would have been possible to enrich for well edited genes, either by selective pull down during the sample preparation or adaptive sampling during ONT sequencing, neither of these were implemented in order to avoid introducing a new source of bias to the method. However, if future experiments using these methods find that the

bias introduced is small or easily corrected. then it may be possible to get more data for the same unit cost of sequencing.

The main limitation of the per-site age method compared with the per-transcript method is that it is constrained to estimating the mean of the transcript ages. Whilst this is sufficient for applications such as half-life estimation, it obscures the richer information found in the per-transcript method (such as in Figure 3.4D) which provide insights into what is causing the change in the mean age. The analyses of the short-read and long-read data in Figure 3.7 showed that the mean gene ages from the short-read data were more accurate. However, this conclusion is drawn from the empirical performance on one dataset, and the sequencing depths of the short-read and long-read were not equivalent. Moreover, the correlation of the per-site and per-transcript models improved when the long-read sequencing depth was increased (Pearson correlation coefficient improved from 0.77 to 0.92).

In the future, simulations of sequencing data could be used to investigate how the errors of the per-site and per-transcript age estimates change with the key variables identified in this Chapter: the sequencing depth, the number of sites per gene and the editing rates of those sites. The approach of simulating data has proved useful for analysing the different models that can be applied to metabolic labelling methods, as detailed in Rummel et al. (2023). A priority of such work should be creating an error metric for the transcript age estimates, $\hat{\tau}_r$, which may provide an improved method to filter out the time-invariant 'zero-edit' reads discussed in Section 3.4.1.

The work presented in Section 3.4.3 highlighted the importance of the upstream data processing pipeline chosen and the impact of 3$^{\text{rd}}$-party software and code. In particular, the discrepancies between the editing information retrieved from JACUSA2 (Piechotta et al., 2021) compared with the combination of IsoQuant (Prjibelski et al., 2022) and Algorithm 1 led to non-trivial differences in the results of the per-transcript method (Figure 3.6D,E). To avoid the biases introduced by JACUSA2's thresholds of minimum number of reads and editing, future efforts may explore using

the SAMtools `mpileup` function (which does not use thresholds), and subsequently repeating the comparisons with the IsoQuant/Algorithm 1 workflow.

The discovery from applying the per-transcript method to the NLambda HEK calibration dataset that many of the 'fit sites' previously identified in Chapter 2 were absent from long-reads was surprising. It was initially assumed that this was largely the result of splicing, since a large proportion of the fit sites were annotated to introns. Due to the difficulty of distinguishing different isoforms from short-read sequencing data and the decision to align all of the sequencing data against a reference genome rather than a transcriptome, discerning the contributions of introns from pre-mRNA vs retained introns was not possible. However, the phenomenon of absent sites persisted even after the list of fit sites was restricted to only those in the 3'UTR - thereby excluding all annotated introns in the reference genome used. Inspection of the 3'UTR regions of some of these genes revealed differences in coverage between the Illumina, PacBio and ONT data. Although the cause of this was not investigated nor quantified across all the genes present, some plausible explanations can be imagined: RNA secondary structure interfering with steps of the library prep (such as the reverse transcriptase), biases of the sequencer technologies for purine enriched sequences, poor performance of the aligners in regions with large amounts of editing, or novel splicing in these regions. In any case, the filtering criteria chosen here (selecting fit sites that appear in the 3'UTRs of genes with at least five 3'UTR sites) could be optimised which would likely increase the number of MLEs that can be calculated from the data whilst reducing the number of noisy transcripts.

Finally, it was demonstrated that mRNA half-lives can be estimated from the age methods detailed in this Chapter. Whilst the three methods presented show good agreement between each other, comparison with established methods such as SLAM-seq (Herzog et al., 2017) are needed to benchmark their performance. Attempts to benchmark against literature values here were unsuccessful due to the difficulty in finding appropriate datasets. Future experiments using SLAM-seq followed by deep long-read sequencing would allow ERA-derived half-lives to be directly compared

with metabolic labelling, and perhaps even enable the integration of 4sU data with ERA estimates. However, 4sU is known to affect the stability of dsRNA (the editing substrate of ADAR1) so validation that this does not interfere with editing may be needed (Gładysz et al., 2019).

Although metabolic labelling was not performed, half-lives were calculated from the calibration experiment by adapting the method of Viegas et al. (2023). This showed some promise and could be refined by optimising the criteria used to define the control set of stable genes used for normalisation. The finding that the Viegas et al. (2023) half-lives were globally higher than the ERA-derived half-lives is similar to a recent comparison of metabolic labelling-derived half-lives with transcriptional halting (ActD) methods (Rummel et al., 2023). If a future comparison between SLAM-seq and ERA shows a strong correlation, then this may support the notion that ActD methods are biased to predicting older half-lives. Compared with both metabolic labelling and transcriptional halting, the ERA-derived half-lives are unique in that they can be estimated directly from single-measurement RNA-seq with no chemical or genetic modification. However, with the current filtering, the half-lives of only around 800 genes can be measured, limiting the breadth of analysis.

The experiments and subsequent analysis in this chapter and the chapter before have focussed on populations of cells treated with Actinomycin D or at steady state. Having now established the fidelity of both per-site and per-transcript mean gene age, in the next Chapter I demonstrate how mean gene ages change in response to stimulation and single-molecule age estimation in single-cells.

Chapter 4

Transcript ages encode changes in gene expression

4.1 Introduction

The previous Chapters have demonstrated that endogenous A-to-I editing can be used to infer the ages of transcripts and pools of transcripts in human cells and that these can measure the change in age after transcriptional halting as well as steady state dynamics. Most biomedical research, however, focuses on the response of cells to perturbations, whether these are external in origin (such as toll-like receptor-antigen binding) or internal (such as transitions in the cell cycle or differentiation). Unlike the demonstrations in Chapter 3 in which transcription was halted or the cells were at rest, most research is concerned with actively transcribing cells and involves determining statistical significance of changes between conditions.

Revisiting the simple model of mRNA turnover presented in Chapter 1 Section 1.3, the major changes in gene expression resulting from a perturbation can be due to either increases or decreases in the transcription rate, or to increases or decreases in the degradation rate. Changes to the transcription rate - rather than the degradation rate - are thought to be the primary mechanism (Rabani et al., 2011). Typical RNA-seq workflows and software packages for analysis (such as DESeq2 (Love et al., 2014)) seek to quantify changes to the 'abundance' of transcripts between two conditions: where abundance is defined as the normalised counts of reads mapping to a particular gene. Normalisation of reads is not a trivial exercise, and the methods

used (such as transcripts per million (TPM), read/fragments per kilobase per million (RPKM/FPKM), spike-ins etc.) see continual development. One major limitation of RNA sequencing is what is known as the RNA composition effect, which can result in the abundance of one gene appearing to decrease between two conditions due to another gene increasing (and *vice versa*) when the true number of transcripts has actually remained unchanged. Whilst the abundance methods are simple to implement and generally reliable, they do not attempt to capture the dynamics of the underlying pool of transcripts.

In this Chapter, I explore the applicability of the Endogenous RNA Age (ERA) methods that I have developed to measure changes in human cells. This involved testing the method in increasingly complex biological systems, with most analysis focussing on the response of primary human monocytes to lipopolysaccharide (LPS) exposure. Monocytes are myeloid cells that make up 2-10% of the white blood cells in the peripheral blood and are important in both innate and adaptive immunity. In response to infection, they traffic out of the blood and become tissue-resident, where they can further differentiate into macrophages and dendritic cells (Shi and Pamer, 2011). A well-studied response in monocytes is to LPS, a component of the outer membrane of Gram-negative bacteria, which is recognised by TLR-4 and triggers signal cascade in a large number of pathways, resulting in a secretory inflammatory phenotype (Guha and Mackman, 2001). This broad transcriptional rewiring presented an ideal setting to test ERA.

This Chapter is structured around several overarching research questions given below:

1. **Can ERA detect promoter-driven changes to gene expression?**
2. **Can ERA identify changes to gene expression due to endogenous regulation?**
3. **Can ERA function in unmodified primary human cells?**

4. **Can ERA determine the ages of genes in single cells?**

4.2 ERA distinguishes between three induced transcriptional programmes in HEK cells

Chapter 3 established that the age methods, now referred to as Endogenous RNA Age (ERA), can accurately measure the ageing of individual RNA transcripts and populations of transcripts. With the eventual goal of studying the changes in gene expression dynamics under endogenous regulation, it was first investigated whether ERA could capture such information for genes under inducible control (using the Tet-On expression system).

Three genes (ATG14, BVES, ACBD7) were selected for testing, based on the presence of many fast edited sites and their per-transcript age performance (full details of transfection methods can be found in Appendix 1). The three stimulation types - ON, OFF and PULSE - were performed as follows (Figure 4.1A):

- ON: Doxycycline (Dox) was added at $t = 0$ hours. Cells were lysed at $t = 0, 0.5, 1, 4$ and 8 hours.
- OFF: Dox was added at 24 hours before the $t = 0$ time. Cells were changed to Dox-free media at $t = 0$ hours and then lysed at $t = 4, 8$ and 16 hours.
- PULSE: Dox was added at $t = 0$ hours. At $t = 4$ hours, ActD was added to halt transcription. At $t = 8$ and 16 hours, the cells were lysed.

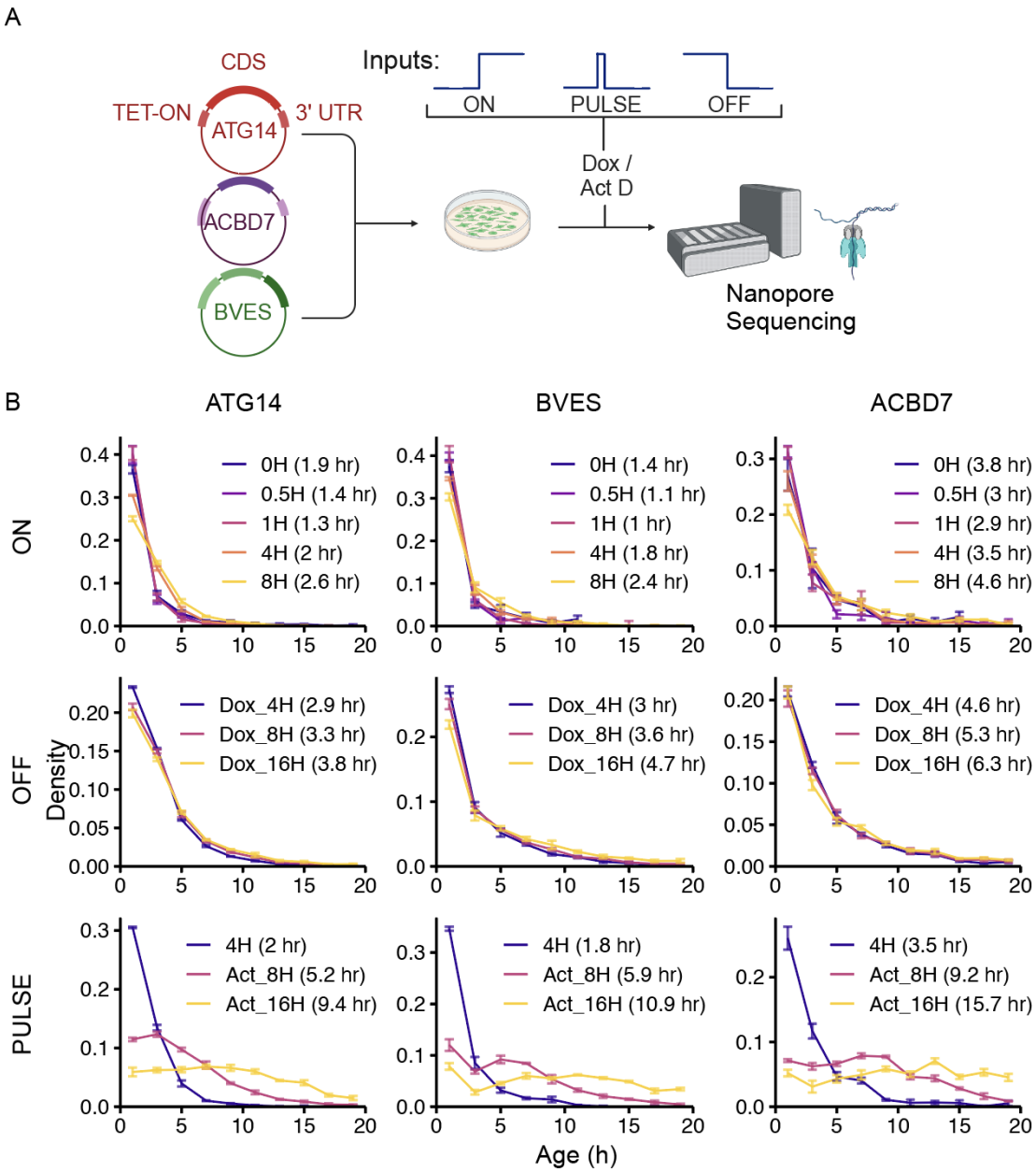

Figure 4.1: **ERA reveals changes in transcript ages for genes under Tet-On control.** (A) Schematic of the experimental set up for the induction of ATG14, BVES, and ACBD7 in HEK cells. 'Inputs' shows a simple representation of the expected gene expression profile over time for each of the inductions: ON (Doxycycline added), OFF (Doxycycline removed)and PULSE (Doxycycline added followed by ActinomycinD). (B) Transcript ages are estimated by maximum likelihood estimation and plotted as density distributions (each normalised to have the same area) for the three genes (columns) in each of the inductions (rows). The specific time point, along with the mean of the transcript age estimates, are given in the figure legends. For the ON legends, the time post-Dox addition is given. For OFF the legends give the time since Dox was removed and for PULSE the legends show the 4h post-Dox condition and then the time since the ActD was added.

Long-read sequencing of the samples was performed on a PromethION24 sequencer over five R10.4.1 flow cells, each yielding approximately 600k mapped reads per replicate (1.8M reads per condition). The transcript MLEs were calculated for each replicate individually and the overall distribution of transcript ages for each condition was calculated by taking the average of the normalised MLE histogram for each condition (Figure 4.1B).

For the ON stimulation (Figure 4.1B, top row) in each of the genes the distribution of transcript ages and the mean age decreased after Dox addition, reflecting the production of new transcripts from the promoter activation. After the 1h time point, the mean age then began to increase, reaching a maximum in the final (8h) time point. The distributions showed a reduction in the proportion of young transcripts (the 0h-2h bin) for the 8h condition and an increase in the proportions of older transcripts. Taking these observations together, ERA detects the initial increase in young transcripts associated with increased transcription and the subsequent increase in age as these transcripts age over the course of the induction.

The mean age for the latter time points was higher than the starting time point and the distribution of the MLEs did not return to that of the initial time point (0h). This was thought to either be due to the system still approaching steady state (at a higher level of transcription) or that the system was instead approaching a different steady state. For a population of transcripts at steady state, the mean age is determined solely by the degradation rate so repeating the experiment but with even later time points may reveal which of the two hypotheses is more likely to explain the results. Of note, the three genes have different starting mean ages, with BVES having the youngest mean age and ACBD7 the oldest. Comparing the change in mean age over the duration of the experiment showed that ACBD7 took the longest to return to or exceed its 0h mean age following the initial decrease after Dox addition. These results are consistent with the theoretical framework of fast turnover genes returning to steady state faster than genes with older mean ages. Put simply, the transcripts of fast turnover genes decay quickly so information on

previous changes to transcription rate are lost faster.

The PULSE stimulation - Dox at 0 hours followed by ActD at 4 hours (Figure 4.1B, bottom row) - exhibited similar behaviour to the calibration experiments, which were also treated with ActD. Further discussion of this can be found in the previous chapter in Section 3.4.

For the OFF stimulation (Figure 4.1B, middle row), the lack of a Dox_0h condition limited the analysis. Naively, it was assumed that this would be the same as the ON 8h condition, however, the difference between the ON 0h and ON 8h conditions meant that this could not be assumed. Nonetheless, there were several phenomena observed in the data. First, when compared with the ON stimulation, for all three genes the mean ages were older and the distributions were wider - which suggested either a decrease in transcription or an increase in transcript stability. Second, the mean age of all three genes increased the more time had passed since the OFF stimulation (the change to Dox-lacking media). However, given that the ON stimulation also produced an increase in mean age after the initial decrease, this suggested that distinguishing between increases and decreases in transcription from mean age alone may be challenging. This was tested by hierarchical clustering of the MLE histograms (Figure 4.2).

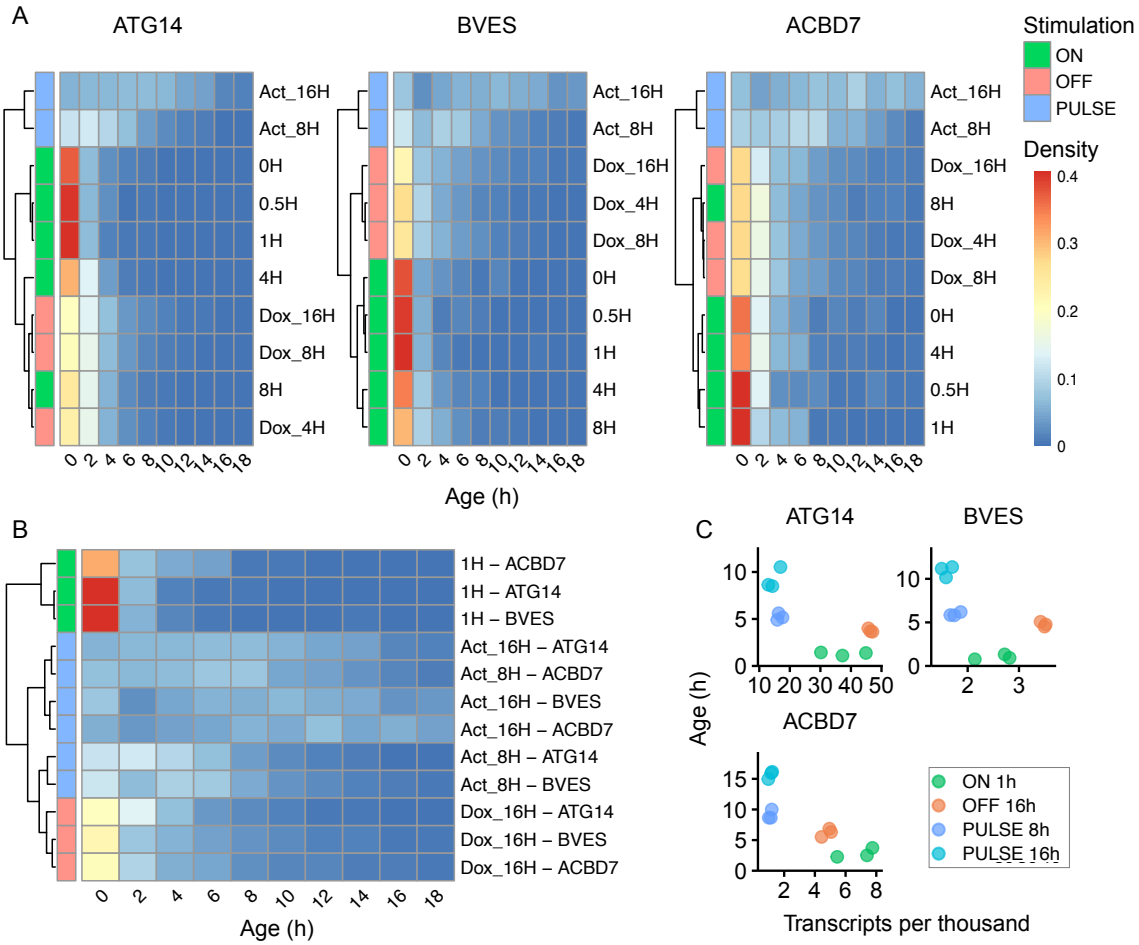

Figure 4.2: **Stimulations of genes under Tet-On control produce patterns in the distributions of transcript ages that are revealed by hierarchical clustering.** (A) For each of the three genes under Tet-On control - ATG14, BVES and ACBD7 - the transcript ages from each of the conditions are converted to histograms and hierarchically clustered. Colours on the left label each row with the corresponding stimulation programme (ON, OFF, PULSE) and the labels on the right give the specific condition. The cells are coloured according to the amount (i.e. density) of transcripts in a given age bin. The ON and OFF stimulations can be resolved - with the exception of the 8h 'ON' and Dox_16h 'OFF' which cluster together for ATG14 and ACBD7. (B) Formatting is the same as in A but one condition from ON, OFF and two from PULSE are selected from each of the three genes and clustered together - revealing that the data cluster first by stimulation type and then by gene. (C) For the same conditions as in B, the mean age is plotted against the abundance (given in transcripts per thousand (TPM times by 1000 for readability)) showing that ON and OFF are more reliable separated by age than abundance.

Clustering all the conditions for each gene in turn (Figure 4.2A) showed that PULSE was clearly distinguished from ON and OFF. For BVES, all of the ON conditions clustered together, as did the OFF conditions. However, for ATG14 and ACBD7,

the ON 8h condition clustered with the Dox_4h (the first OFF condition) indicating that the distribution of transcript ages were similar between these two conditions. In general, the OFF stimulations (and in particular, the Dox_4h condition) did not exhibit as dramatic a change as ON or OFF - suggesting that the impact of the media change was slow to take effect, especially in contrast to the halting of transcription caused by ActD in the PULSE stimulation.

To better compare the three stimulations, I selected only the conditions from the ON and OFF stimulations that were furthest from the 0h in terms of mean age: the ON 1h and the OFF 16h (Dox_16h). I clustered these two conditions along with the two PULSE conditions for all three genes at once (Figure 4.2B). This plot revealed that the age distributions first clustered by the stimulation type and then by gene - demonstrating the consistency of the response of the three genes to each of the stimulations.

Finally, it was hypothesised that integrating measures of transcript abundance (a typical output of RNA-seq) with the age estimates may make it easier to resolve between these two behaviours. Selecting the same conditions as Figure 4.2B, Figure 4.2C shows the abundance (in transcripts per thousand) on the $x-$axis vs the mean age on the $y-$axis. Whilst there was no consistent difference in the abundance for the ON vs the OFF samples across the three genes, the ON 1h was consistently younger than the OFF 16h - suggesting that age rather than abundance is the more useful metric for classifying the stimulations here. In addition, the PULSE samples showed little difference in abundance, yet were trivially discernible by age. It may be the case that selecting a different normalisation method for abundance other than transcripts per million may be more appropriate for the circumstances of this experiment (that is, changes to the expression of transgenes) yet the transcript age histograms have the advantage of not needing to be normalised by library size at all.

Taken together, the results from the stimulation experiment demonstrated that ERA

detects both activation and inhibition of gene expression for three transgenes expressed under the Tet-On system. We next explored whether ERA could detect age changes of genes under endogenous regulatory control in HEK cells.

4.3 ERA discovers changes to transcript ages in HEK cells responding to heat shock

Having screened several different stimulations of HEK293 cells, including treatment with TNF, Forskolin, Lithium Chloride and DMOG (to induce hypoxia), heat shock was selected as the perturbation most likely to produce changes to the expression of well edited genes. The heat shock was carried out by collaborator, Ali, and performed by moving the HEK cells from the standard culturing temperature of 37°C to an incubator at 42°C followed by lysis at 0h (before heat shock), 1h and 4h. Additionally, there was a control condition in which the cells were moved to another incubator at 37°C and lysed after 4 hours. The 12 samples were pooled and sequenced on a PromethION24 over nine R10.4.1 flow cells, yielding an average of 45.8M reads per sample (137M reads per condition in triplicate).

The data were first analysed with DESeq2 (Love et al., 2014) to identify differentially expressed genes from the IsoQuant (Prjibelski et al., 2022) abundance output files. For the thresholds of $|\log_2(\text{FC})| \geq 1$ and adjusted p value (adj.p, by the Benjamini-Hochberg method) < 0.05, the analysis of the 1h-0h contrast identified 14 genes as differentially expressed (12 upregulated, 2 downregulated), for the 4h-0h contrast 290 were identified (132 upregulated, 158 downregulated) and for the 4h-1h contrast, 165 genes were differential (74 up, 91 down) (Figure 4.3A, top row). Only one gene was called as differentially expressed for the 4h_control vs 0h condition, suggesting that time alone had minimal impact and that the changes seen were due to the heat shock. Pathway enrichment analysis (Gene Ontology, Biological Process) of the same contrasts showed terms enriched for protein folding and heat response, with the 4h having more genes enriched in similar terms compared with the 1h (Figure

4.3B,C).

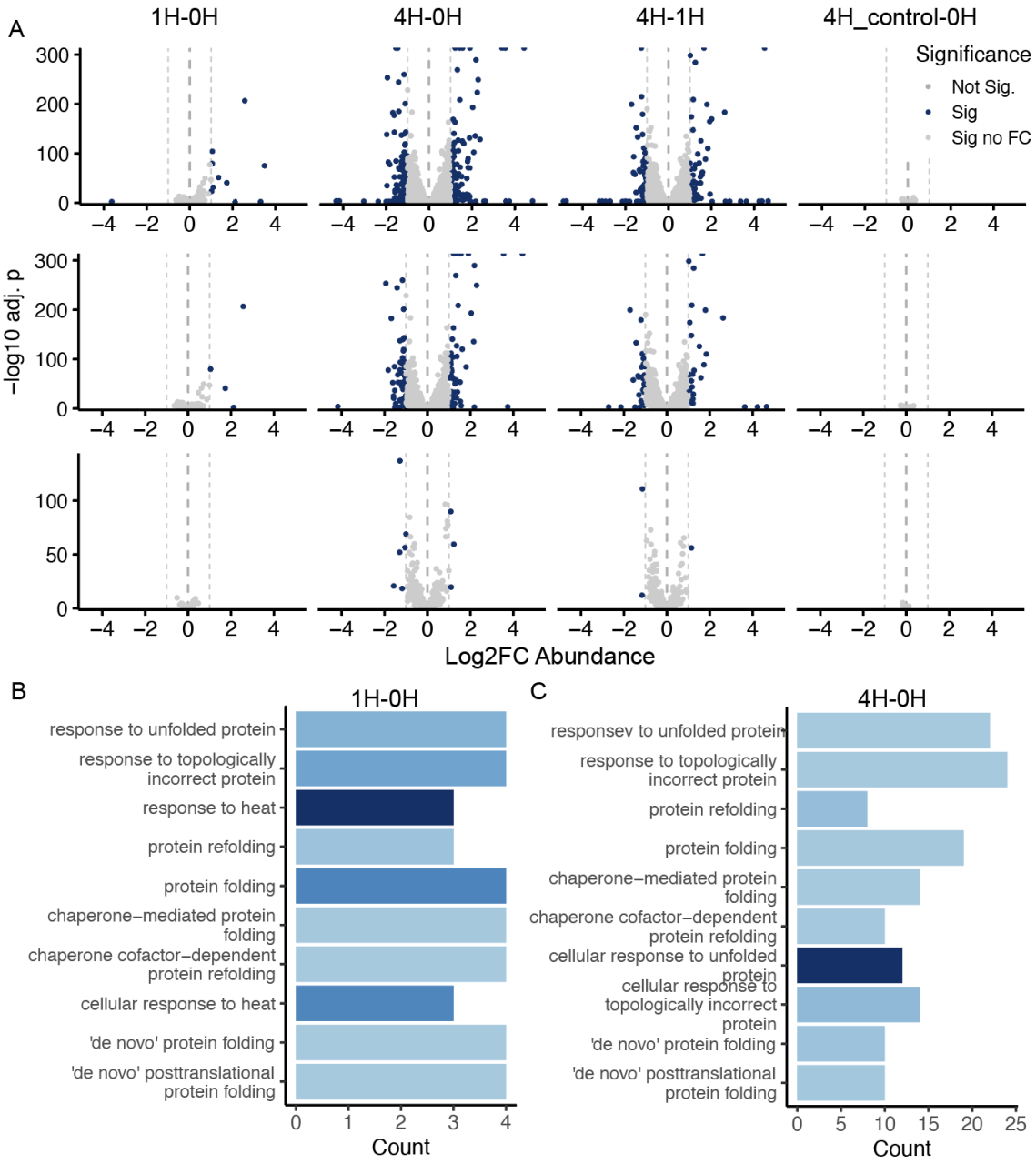

Figure 4.3: **Differential expression analysis reveals characteristic gene expression signatures in HEK cells exposed to heat shock.** (A) Volcano plots showing the differentially expressed genes (DEGs) for each of the two experimental contrasts (the 1h and 4h - relative to the 0h), the 4h-1h contrast and the 4h_control-0h contrast. Top row: DEGs for all genes tested; middle row shows the same but subset to genes that have fit sites; bottom row is the same but subset to genes that have 5 or more 3'UTR fit sites. (B) Gene Ontology Biological Process (GO BP) terms enriched among all differentially expressed genes ($|\log_2(\text{FC})| \geq 1$, adjusted p value < 0.05) 1h and 4h post-heat shock, relative to the 0h condition. Bars are coloured by adjusted p value with with lighter being smaller. All bars have p $<$ 0.05. Enriched terms identify protein refolding activity that is characteristic of heat shock.

Subsetting the results above to only genes that contain fit sites ($n = 949$ genes) left no differentially expressed genes in the 1h-0h contrast, 9 in the 4h-0h contrast (3 up, 6 down) and 3 in the 4h-1h (1 up, 2 down) (Figure 4.3A, bottom row). With so few genes, no pathway enrichment was detected in any contrast. The fact that most of the genes that are identified by DESeq2 as differentially expressed after heat shock did not contain fit sites - either because these genes were not expressed highly enough during the calibration experiment or because they contain no editing sites - suggested that ERA may be unable to detect many of the hallmark responses from the stimulation. Since the list of fit sites used was restricted to those in the 3'UTR of genes with at least five 3'UTR sites, the analysis was re-run for all genes that contain at least one fit site anywhere on the gene sequence (Figure 4.3A, middle row). This retained approximately half of the differentially expressed genes identified prior to filtering, indicating that improvements to the filtering criteria applied to fit sites would substantially improve the number of genes that can be analysed by ERA. This would be especially useful for pathway analysis, where the size of the background 'universe' of genes in which pathway enrichment is tested can significantly change the results obtained.

The ages of transcripts were estimated as described in Chapter 3, Section 3.4 and Methods, yielding an average of 327,936 MLEs per condition. Unlike in the HEK induction experiment described previously in Section 4.2 where the identity of the genes of interest was known, here a new method was needed to identify genes with statistically significant changes in their transcript ages. For this I developed a metric termed 'differential age', which provides an analogous output to the 'differential expression' of DESeq2. For each gene, it calculates a $\log_2$ fold change ($\log_2$ FC), a p-value and an adjusted p-value. The $\log_2$ FC is the $\log_2$ transformed mean of the estimated transcript ages of condition 2 divided by the same in condition 1. Based on the previously observed changes in mean age in the induction experiment (Section 4.2), I selected a threshold of $|\log_2(\text{FC})| \geq 0.5$ for the effect size. The p-value is determined by a Mann-Whitney U test on the MLE values between the two

conditions and the false discovery rate adjustment was performed by the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

The results of the differential age method are summarised below in Table 4.1 and Figure 4.4.

Contrast	**nGenes**	**Sig Genes**	**Increased**	**Decreased**
1H-0H	849	1	0	1
4H-0H	845	69	62	7
4H-1H	847	47	44	3

Table 4.1: **Differential age analysis of edited genes in response to heat shock**. 'Contrast' denotes the two conditions tested for differentially aged genes. 'nGenes' denotes the number of genes present in the given contrast. 'Sig Genes' denotes the number of genes that have an adjusted p value < 0.05 (Benjamini-Hochberg corrected) and a $|\log_2 \mathrm{FC}| \geq 0.5$. 'Increased' denotes the subset of 'Sig Genes' that increased in age whilst 'Decreased' denotes those that had a reduction in age. A reduction in age is typically produced by a recent increase in transcription whilst an increased age is associated with either a recent decrease in transcription or a distant increase in transcription (Section 4.2).

Testing the differentially aged genes (Table 4.1, Figure 4.4A) for enriched GO BP terms identified no significant results, likely as a result of the low number of genes in each contrast. I examined genes that were called as differential in age but not in abundance . For the 1h-0h contrast, only ANAPC16 was identified which was called as significantly decreasing in age from 11.8h to 7.72h ($\log_2$FC(Age) = -0.61, adj.p value = 0.018) but not significant from differential expression analysis ($\log_2$FC(Abundance) = 0.07, adj.p value = 0.22) (Figure 4.4B). ANAPC16 is a one of the subunits of the 1.5MDa anaphase-promoting complex/cyclosome, a critical E3 ubiquitin ligase for cell cycle progression, particularly for metaphase to anaphase transition (Chang et al., 2014). Inspection of the transcript age distributions reveals a sharp increase in young transcripts in the 1h condition, suggesting a burst of transcription (Figure 4.4B). The slight but not significant increase in abundance detected by DESeq2 suggests that differential age - by segmenting the transcripts by age - has higher sensitivity for small but recent changes in transcription. ANAPC16 also has

an old mean age both in the 0h condition (11.8h) and the 4h control (10.9h), and this slow turnover further increases the sensitivity to changes in transcription. Inspecting the distribution of estimated transcript ages from the 4h condition showed that the fraction of young transcripts had decreased, indicating that the burst of transcription in the 1h was transient (Figure 4.4B). Although heat shock has been documented as delaying mitosis, an early increase in transcription of ANAPC16 or other components of the ACP/C has not previously been reported (Kakihana et al., 2019; Ota et al., 2023).

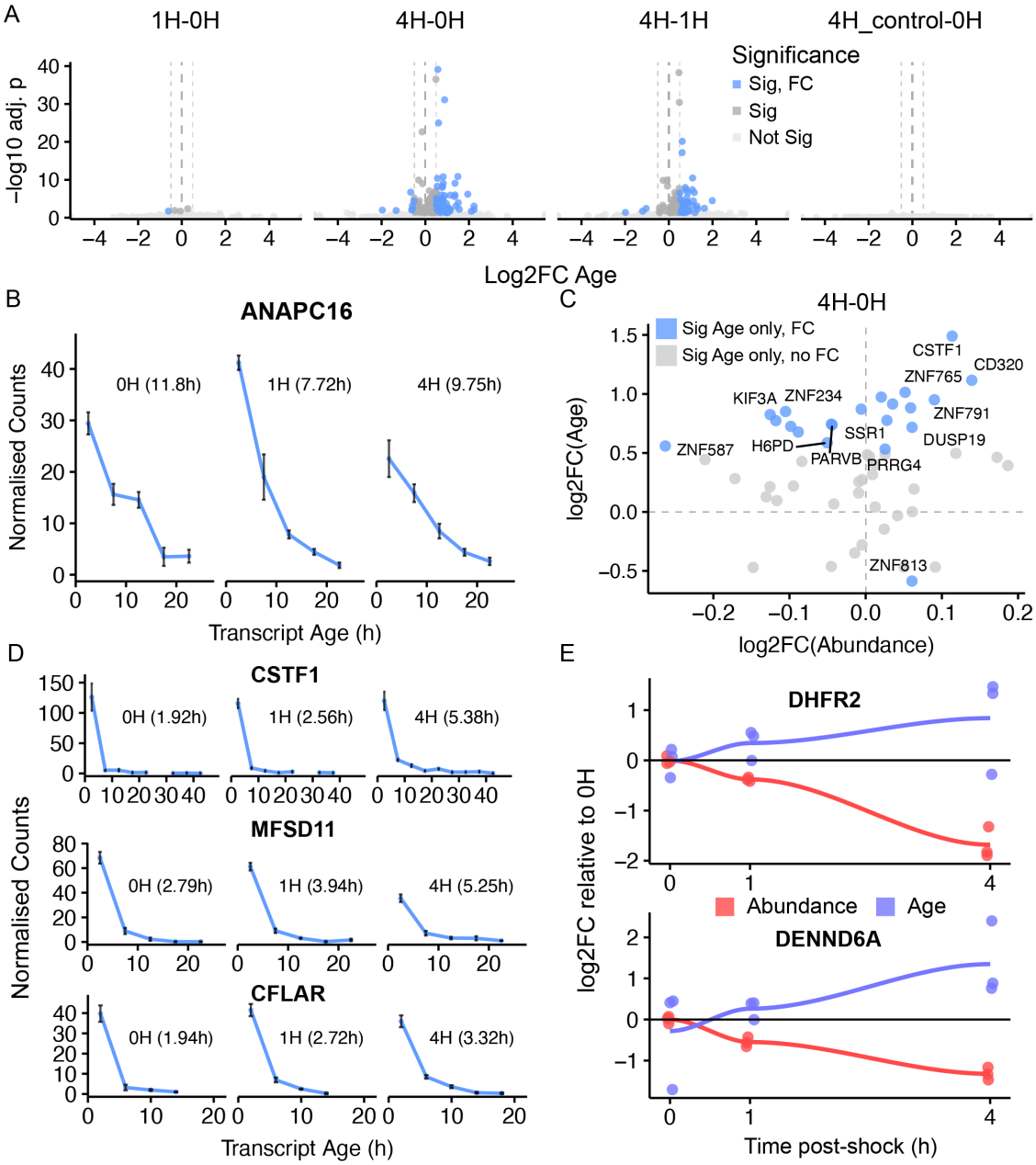

Figure 4.4: **Differential age analysis identifies genes responding to heat shock in HEK cells** (A) Volcano plots showing the differentially aged genes ($|\log_2(\text{FC})| \geq 0.5$ and adjusted p value < 0.05). (B) Distributions of the transcript ages are shown for ANAPC16 - the sole gene identified as being differentially aged but not differentially expressed 1 hour post-heat shock - with an increase in young transcripts and a decrease in old transcripts. The mean age is given at the top of the plots. (C) Differential age vs differential abundance is shown for genes 4 hours post-heat shock. (D) Transcript age distributions showing a potential stabilisation of older transcripts (CSTF1), a decrease in transcription (MFSD11) or a combination of both (CFLAR). (E) The two genes that are significant for both age and expression - DHFR2 and DENND6A - are shown with the $\log_2$FC relative to the 0h values for abundance (red) and age (blue) respectively, showing age increasing as abundance decreases.

Repeating the analysis of differentially aged but not differentially expressed genes for the 4h-0h condition yielded 21 genes (20 increasing in age, 1 decreasing) (Figure 4.4C). The one decreasing gene was ZNF813, a zinc finger protein of unknown function, which exhibited a significant and sustained decrease in age in the 1h and 4h time points ($\log_2$FC(Age) = -0.59 for both time points compared with the 0h), driven by an increase in the proportion of young transcripts. The 20 genes that increased in age but had no significant change in abundance were generally either zinc-finger proteins (which are highly represented in the calibration data), proteins involved in stress response (CFLAR, MCUR1, DUSP19), and vesicle trafficking or metabolite transport (H6PD, CD320). The distribution of estimated transcript ages in several of these genes suggested that transcript stabilisation (i.e. a decrease in the degradation rate) may be responsible for the increased ages (CSTF1 (Figure 4.4D top row), CD320, PRRG4), whereas for others a decrease in transcription rate seemed more likely (ZNF765, MFSD11 (Figure 4.4D middle row), PLEKHA2, SSR1, PARVB, AP1S3, DUSP19, TANG02, H6PD). For others, a combination of both seemed plausible (SLC16A10, ZNF791, ZNF562, ZNF234, KIF3A, CFLAR (Figure 4.4D bottom row), MCUR1, ZNF587). For some, the data suggested that the transcripts had already been stabilised in the 1h time point and then the transcription rate proceeded to drop in the 4h time point (CFLAR, MCUR1).

The only genes that passed the significance and effect size thresholds for both differential age and differential abundance were DENND6A and DHFR2 in the 4h-0h contrast (Figure 4.4E). Both of these genes appear to be downregulated, as evidenced by a decrease in expression and an increase in age. DHFR2 is from the same family as DHFR, a critical component of folate metabolism which is important for DNA synthesis and can be bound by the heat shock protein HSPA1A (also known as Hsp72) under cellular stress (Musch et al., 2004). Its downregulation here (and a large upregulation of HSPA1A ($\log_2$FC(Abundance) = 2.28)) may indicate the cells de-prioritising replication in response to the heat shock. Interestingly, HSPA1A had 12 fit sites in the calibration data yet since these were exonic and not in the 3'UTR

it was not present in the list of genes tested for differential age. DENND6A was recently implicated in lysosomal trafficking and its knock out in human cell lines reduced autophagy (Kumar et al., 2024). Whilst autophagy and lysosomal activity are canonically important in stress response, DESeq2 did not identify large $\log_2$FC changes in either the 1h-0h or 4h-0h contrasts in the hallmark genes TFEB, ATG5, ATG7, BCEN1 or SQTM1, leaving the downregulation of DENND6A without a clear explanation.

4.4 Response of primary human monocytes to LPS treatment

Monocytes are an attractive cell type to investigate with ERA, since they are easy to obtain by isolation from peripheral blood and are highly responsive to stimuli. Seeking to induce a large transcriptional change (to increase the probability of well-edited genes being differentially regulated, which was one of the limiting factors of the previous heat shock experiment in HEK293 cells), monocytes were treated with lipopolysaccharide (LPS). LPS is a component of the outer membrane of Gram-negative bacteria - and is well studied in perturbations of myeloid lineage cells, in which it is known to cause broad transcriptional reprogramming across hundreds of inflammation-associated genes (Guha and Mackman, 2001; Rabani et al., 2011).

Monocytes were isolated from the blood of a sole healthy male based in the UK (with thanks to Gabrielle Chappell for sharing her protocol and performing much of the isolation protocol) and plated 24 hours prior to stimulation. At $t = 0$h, LPS was added and cells were lysed in triplicate at $t = 0$h (prior to LPS addition), 1h, 2h, 4h and 6h. A 6h control condition was also lysed to which no LPS was added - yielding a total of 18 samples. These samples were pooled and sequenced on a PromethION24 using 18 R10.4.1 flow cells, yielding roughly 70M reads per sample and generating over 1.2B long-reads. A similar analysis workflow as described above for the HEK heat shock experiment was followed for the monocytes.

Differential expression analysis was performed using DESeq2 on the abundance data, the results of which are summarised in Table 4.2 below and Figure 4.5. The number of genes passing both the $\log_2$FC(Abundance) and adjusted p value thresholds was over ten-times higher than in the HEK heat shock experiment ($n = 290$ differentially expressed genes 4 hours post-heat shock compared with $n = 3{,}538$ for the monocytes 4 hours post-LPS), validating the assumption that the LPS stimulation would yield a large transcriptional change in the monocytes (Figure 4.5A).

Contrast	**nGenes**	**Sig Genes**	**Increased**	**Decreased**
6H_control-0H	14043	149	76	73
1H-0H	16428	1488	868	620
2H-0H	16786	2714	1650	1064
4H-0H	17154	3538	2052	1486
6H-0H	17154	3745	2151	1594

Table 4.2: **Differential gene expression analysis with DESeq2 for human monocytes treated with LPS.** The table presents data for each of the conditions relative to the 0h, as well as the control (6h_control) condition. 'nGenes' denotes the total number of genes analysed, 'Sig Genes' represents the number of significantly expressed genes ($|\log_2(\text{FC(Abundance)})| \geq 1$ and adjusted $p < 0.05$), 'Increased' and 'Decreased' indicate the numbers of genes showing significantly increased or decreased expression, respectively.

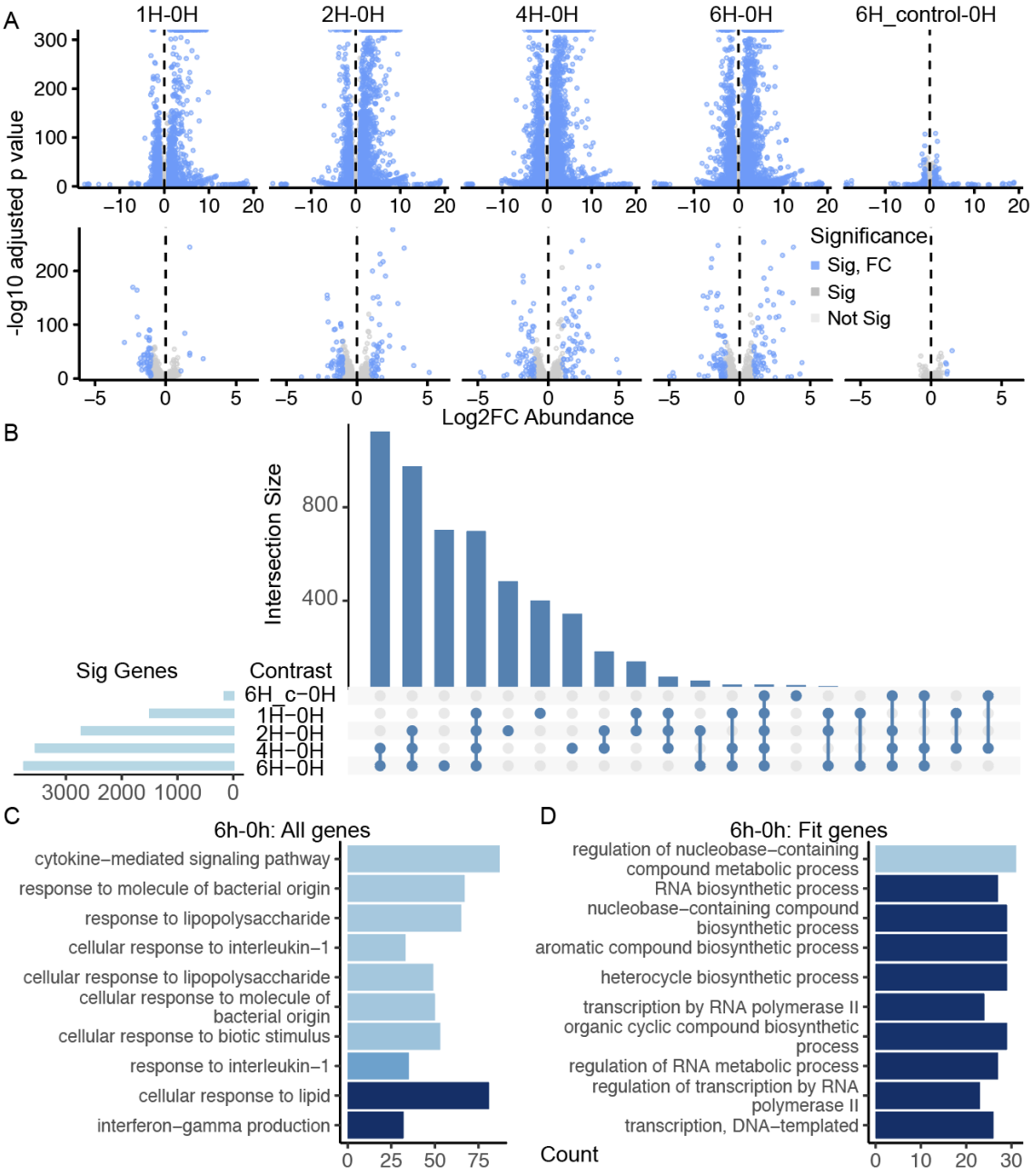

Figure 4.5: **Differential expression analysis of time series long-read sequencing of primary human monocytes identifies canonical regulatory processes.** (A) Top row: Volcano plots showing that DESeq2 identifies significant changes in expression for hundreds of genes in response to LPS treatment relative to the 0-hour baseline. Bottom row: same as top row but filtered *post-hoc* to well-edited genes in the calibration data. (B). The overlaps between the sets of differentially expressed genes between the different comparisons is visualised as an UpSet plot. The point matrix shows which intersection corresponds to the bar directly above it ('Intersection size'). The 'Sig Genes' in light blue shows the total number of significant ($|\log_2(\text{FC})| \geq 1$, adjusted p-value by Benjamini-Hochberg method < 0.05) genes in each set. (C & D) Enrichment of significant genes for Gene Ontology Biological Process (GO BP) terms for all significant genes (C) and those that contain well-edited sites (D). Bars are coloured by decreasing adjusted p-value from dark to light, all bars shown are significant.

As a general trend, the number of differentially expressed genes - relative to the 0h condition - increased over the duration of the experiment. Of the 5 contrasts tested, the 6h-0h comparison had the largest number of differentially expressed genes, closely followed by the 4h-0h (Figure 4.5B). The 6h_control-0h contrast had very few differentially expressed genes - as may be expected - and most of these were unique to that contrast. Most of the differently expressed genes in the 6h condition were shared with the 4h and 2h conditions and not found solely in the 6h. This suggests that a large fraction of the genes, once differentially expressed, remain differentially expressed for the duration of the experiment - as evidenced by the 6h-4h-2h intersection being larger than the 6h-2h intersection (Figure 4.5B).

Enrichment analysis was then undertaken (Table 4.3) on GO BP terms for each time point (1h-0h (Figure 4.5C), 2h-0h, 4h-0h, and 6h-0h) and the top 50 (ranked by adjusted p-value) terms were used to compare the different contrasts. 19 pathways were significantly enriched in all of the contrasts, indicating a strong and durable core transcriptional response. Enriched pathway terms were associated with response to molecule of bacterial origin, cytokine-mediated signalling pathway and cell-cell adhesion - suggesting that a core innate immune response to LPS is maintained over time. Pathways that were only shared in the early contrasts (1h-0h and 2h-0h) were related to acute inflammatory response, including cellular responses to interleukin-1, chemokine-mediated signalling and interferon-gamma production. The later time points (4h-0h and 6h-0h) uniquely enriched pathways for immune regulation and migration, such as T cell activation, response to interferon-gamma and chemotaxis, all of which agree with the canonical ordering of changes in monocytes stimulated with LPS. The 6h_control-0h contrast, despite the far fewer differential expressed genes, was enriched for terms related to response to metal ion and chromatin remodelling, indicating a potential reaction to metal ion availability in the culture media. Together, these results indicate a sweeping remodelling of the monocyte transcriptome in response to LPS.

Contrast	Enriched pathways	Significant	Unique	Shared
6H_control-0H	1966	37	24	13
1H-0H	4821	458	24	434
2H-0H	5266	848	57	791
4H-0H	5423	1369	178	1191
6H-0H	5462	1286	144	1142

Table 4.3: **Summary of GO BP enrichment analysis for differentially expressed genes in human monocytes treated with LPS.** 'Enriched pathways' gives the number of pathways that were enriched, 'Significant' gives the number of these for which the enrichment was significant. 'Unique' gives the number of pathways that were significantly enriched in only that contrast and 'Shared' gives the number that were found in at least one other contrast.

The above analysis was repeated but with the data subset only to well-edited genes (Table 4.4). This produced a much smaller subset of differentially expressed genes ($n = 135$ genes for the monocytes 4 hours post-LPS) but far more than observed for the HEK heat shock experiment ($n = 9$ genes 4 hours post-heat shock). Notably, well-edited genes appeared to be strongly skewed towards a decrease in expression in the 1h-0h contrast (Figure 4.5A bottom row), compared to the set of all genes in Table 4.2. Pathway analysis of the differentially expressed well-edited genes produced a very different signature to the analysis of all the genes. No significantly enriched pathways were shared across all of the time points, early pathways (1h and 2h) were strongly enriched for metabolic or biosynthetic processes and transcription regulation ($n = 34$ significant terms), whereas old pathways (4h and 6h) were enriched for morphogenesis ($n = 3$ significant terms). Neither the 2h nor the 6h time points had any enriched pathways, suggesting that the early changes to expression of metabolic genes had already occurred by the 1h time point (Figure 4.5D) and similarly that the changes in morphogenesis genes occurred between 2 hours and 4 hours post-LPS.

Contrast	nGenes	Sig and FC	Increased	Decreased
6H_control-0H	842	5	5	0
1H-0H	875	54	8	46
2H-0H	879	107	54	53
4H-0H	883	135	70	65
6H-0H	883	150	76	74

Table 4.4: **Differentially expressed genes that are also well-edited in human monocytes treated with LPS.** 'nGenes' gives the number of genes tested in each contrast that also had five or more 3'UTR editing sites. 'Sig and FC' gives the number of genes passing both the significance and fold change (FC) thresholds ($|\log_2(\text{FC})| \geq 1$ and adjusted $p < 0.05$), 'Increased' and 'Decreased' indicate the numbers of edited genes showing increased or decreased abundance, respectively.

Having analysed the changes in the transcript abundance from standard differential expression analysis, the gene ages were analysed using ERA.

4.4.1 ERA discovers differentially aged genes in primary human monocytes treated with LPS

The ages of all transcripts were estimated by maximum likelihood estimation for all samples as described previously. This yielded 1,416,351 MLEs over the whole dataset and an average of 236,059 per condition ($n = 805$ genes on average per condition, out of the possible 949 genes in the calibration dataset), slightly fewer than for the HEK heat shock experiment.

Differential age analysis was performed on each condition relative to the 0h time point with the results shown in Figure 4.6A as volcano plots and in Table 4.5. Fewer genes were identified as differentially aged than as differentially expressed, even when the differential expression results were subset to only well-edited genes (Table 4.4). Whereas for the differential expression results, in which the number of differentially expressed genes increased in successive time points, for the differentially aged genes the 6h contrast had the greatest number of significant results followed by the 2h, 1h

and finally 4h. Analysis of the intersections between the differentially aged genes in each of the contrasts found that most of the genes were only differential in a single condition and just 2 genes (SLC30A4, TMEM120B) were differentially aged in all of the contrasts.

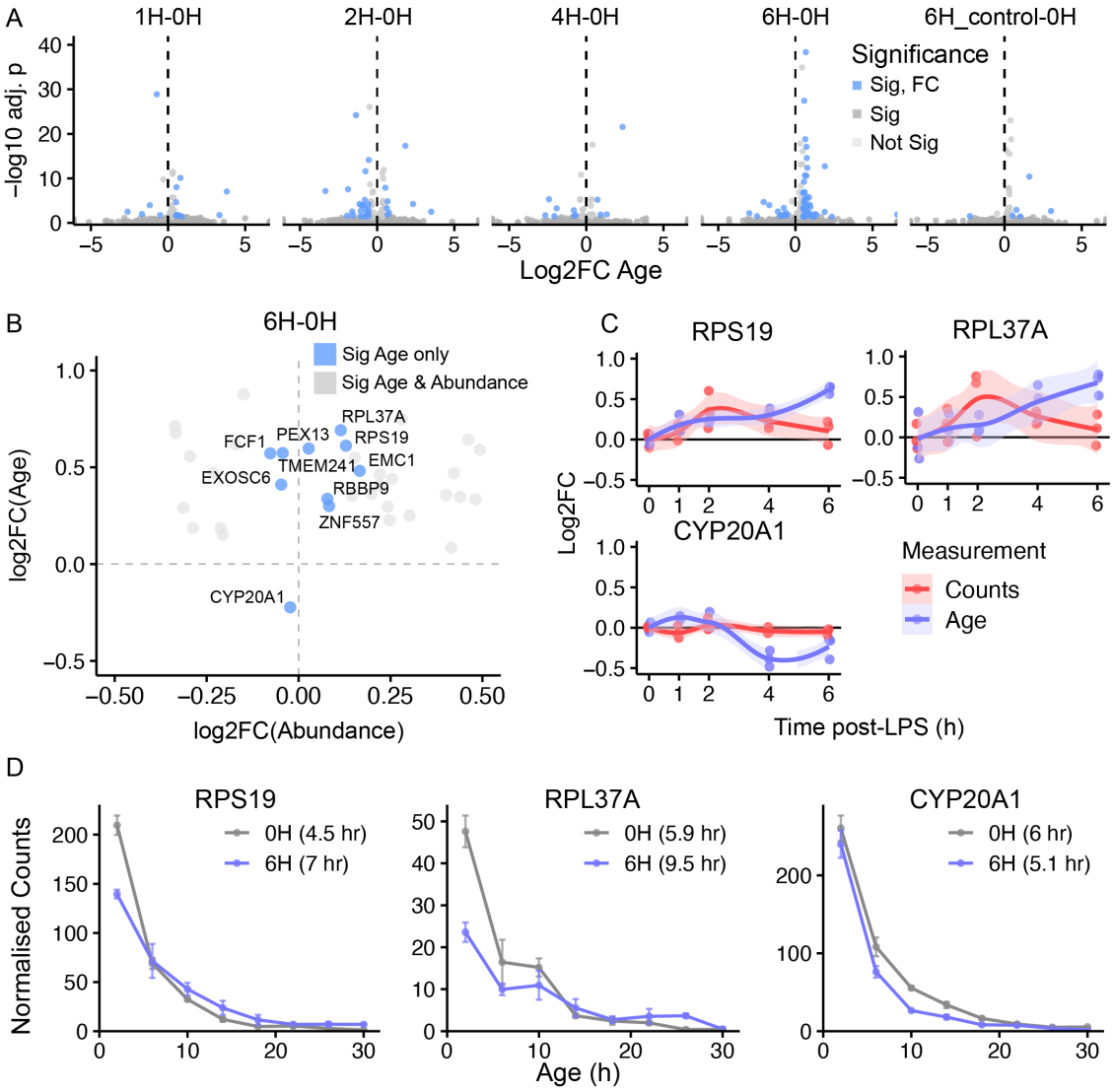

Figure 4.6: **ERA identifies differentially aged genes in human monocytes responding to LPS.** (A) Volcano plots showing the differentially aged genes ($|\log_2(\text{FC(Age)})| \geq 0.5$ and adjusted p-value < 0.05) in response to LPS treatment relative to the 0-hour baseline. (B) Comparison of of the differential aged vs differentially expressed genes 6 hours post-LPS treatment. (C) Trajectory plots that show the change in abundance (red) and age (blue) as $\log_2$(FC) relative to the level prior to LPS addition (0h). RPS19, RPL37A and CYP20A1 were identified from pane B as differentially aged but not differentially expressed 6 hours post-LPS. (D) Distributions of transcript ages for the three genes from C in the 0h (grey) and 6h (blue) conditions. A bin width of 4 hours was used and error bars indicate the standard error of the mean for each bin across the three replicates per condition.

Contrast	**nGenes**	**Significant**	**Sig & FC**	**Increased**	**Decreased**
6H_control-0H	782	28	6	5	1
1H-0H	772	41	16	11	5
2H-0H	775	61	31	8	23
4H-0H	773	33	11	4	7
6H-0H	769	96	59	47	12

Table 4.5: **Differential age analysis of monocytes responding to LPS.** 'Contrast' contains the two conditions tested for differentially aged genes. 'nGenes' denotes the number of genes present in the given contrast. 'Significant' gives the number of genes that have an adjusted p value < 0.05 (Benjamini-Hochberg corrected). 'Sig & FC' are those in 'Significant' that also have a $|\log_2 \text{FC}| \geq 0.5$. 'Increased' indicates Sig & FC Genes that increased in age whilst 'Decreased' shows genes that had a reduction in age. A reduction in age is typically produced by a recent increase in transcription whilst an increased age is associated with either a recent decrease in transcription or a less recent increase in transcription.

ERA's ability to identify transiently expressed genes was then tested by analysing genes that were differentially aged 6 hours post-LPS but not differentially expressed (Figure 4.6B). This identified 10 significant genes, 5 of which had a $|\log_2(\text{AgeFC})| \geq 0.5$, with 9 out of the 10 displaying an increase in age. To visualise the change in abundance and age over the experiment, the $\log_2$FC for each was calculated relative to the mean value in the 0h condition and plotted in what I term 'trajectory plots' (Figure 4.6C). For the two genes (RPS19 and RPL37A, both ribosomal proteins) that showed the largest increase in age in the 6h-0h contrast, both appeared to have a transient increase in abundance that peaked at 2h post-LPS but had returned towards the baseline by the 6h time point. The single gene that decreased in age in the 6h contrast - albeit by a small amount ($\log_2(\text{AgeFC}) = -0.22$) - was a cytochrome p450 gene (CYP20A1). CYP20A1 is the last human cytochrome p450 family gene for which the biological substrate and catalytic function is not known, even though the gene has orthologs in all other sequenced vertebrates (Brun et al., 2021). The transcripts of CYP20A1 did not display any significant changes in abundance in any of the time points, suggesting that the observed decrease in mean age may be

due to either destabilisation of old transcripts, a recent increase in transcription or a combination of the two. Inspecting the distribution of transcript ages for CYP20A1 (Figure 4.6D) showed a decrease in the proportion of older transcripts, suggesting a possible destabilisation event, although the significance of this was not determined. Interestingly, CYP20A1 is the only member of the cytochrome p450 superfamily present in the calibration dataset and is in fact the most heavily edited gene, harbouring a remarkable 312 3'UTR editing sites (CYP4F26P and CYP2D7 feature 230 and 15 intronic editing sites, respectively, but are excluded from analysis since only 3'UTR sites are considered). The discovery of such a high level of editing offers a new avenue through which to investigate the function of this orphan gene, perhaps via a phylogenetic analysis of whether its orthologs also possess high levels of editing sites relative to the other p450 family members in their species. Given that many editing sites are present in *Alu*-repeats and that these are uncommon outside of primates, it would be interesting to see if the CYP20A1 orthologs in lower vertebrates exhibit different spatial or developmental expression patterns.

Having discovered in the HEK heat shock analysis that ERA was sensitive to small recent changes in transcription rates, I next analysed the 1h-0h contrast to investigate the immediate changes to LPS stimulation (Figure 4.7A). Of the 16 genes identified as significant by differential age but not abundance, von Hippel-Lindau (VHL) was the only gene to decrease in age and this was due to an increase of young transcripts, indicating a recent transcription increase. Inspecting two of the genes with the largest increase in age in this contrast - HEATR3 and EIF2AK2 (also known as protein kinase R (PKR)) - showed a decrease in young transcripts and an increase in transcripts of other ages. HEATR3 is involved in ribosome biogenesis whilst EIF2AK2 is a central regulator of multiple cellular process including transcription, translation and proliferation and is directly directly implicated in suppression of Type I Interferon response by ADAR1 editing(Gal-Ben-Ari et al., 2019; de Reuver et al., 2022; Arakawa et al., 2025). An initial decrease in transcription of these two genes in response to LPS has not been previously reported and their fates

diverge after the 1h time point, with EIF2AK2 massively increasing in abundance (Figure 4.7B,C) whilst HEATR3 decreases. Comparing the 4h and 6h transcript age distributions for EIF2AK2 shows the density of estimated transcript ages shifting to the right in the 6h condition, indicating that ERA was recovering the increase in transcription that occurred 4 hours earlier in the 2h condition (Figure 4.7C).

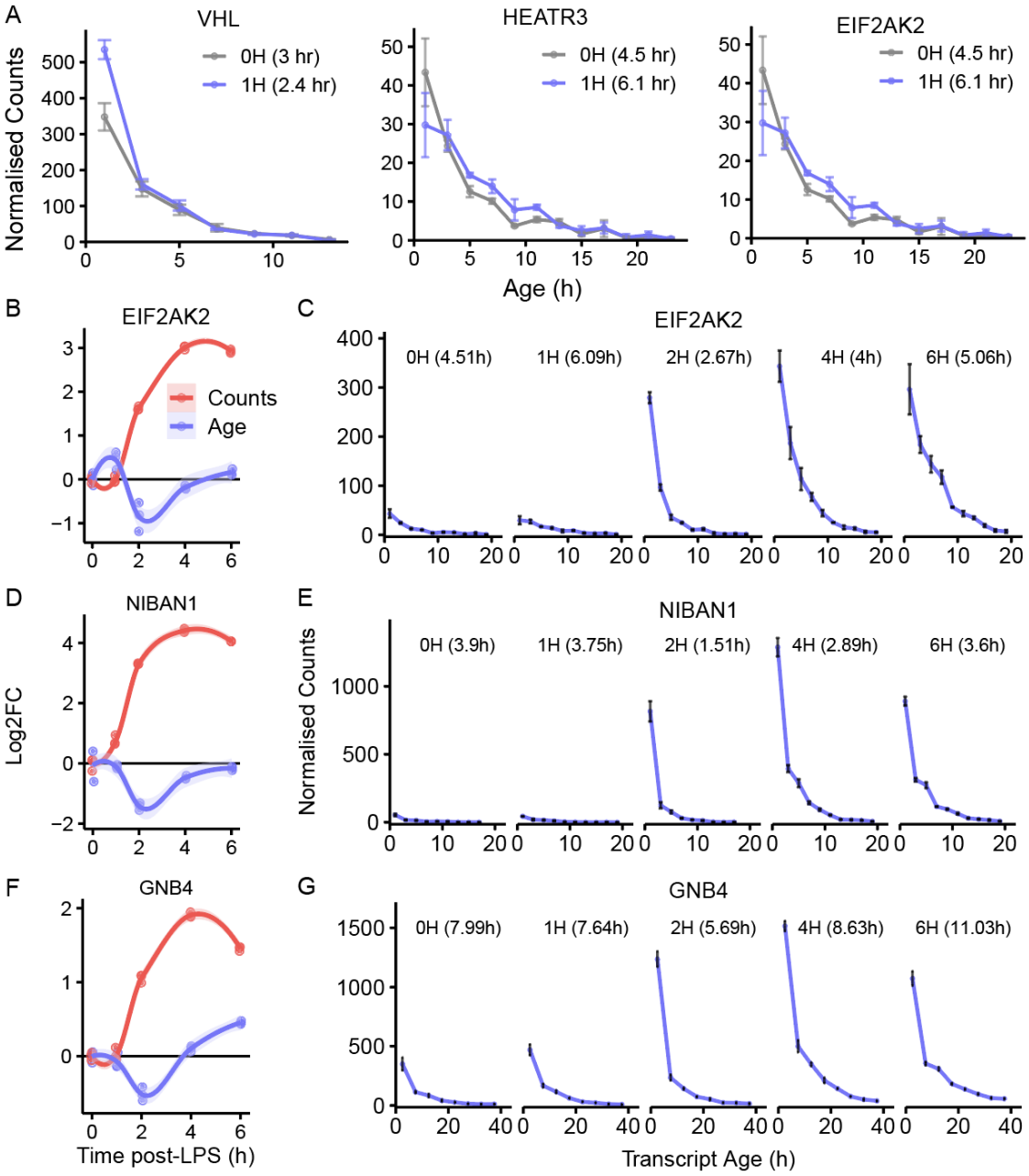

Figure 4.7: **Transcript ages encode the history of past changes in gene expression.** (A) The distributions of transcript ages for three genes (VHL, HEATR3 and EIF2AK2) identified as differentially aged but not differentially expressed 1 hour post-LPS stimulation are shown. A bin width of 2 hours was used to generate the distribution and the mean age of each gene is given in the plot legend for the 0h (grey) and 1h (blue) conditions. (B, D, F) Trajectory plots for three genes that undergo large upregulation, showing the $\log_2$FC in normalised counts (red) and age (blue) relative to the mean of the level in the 0h. The shaded area shows the 95% confidence interval. (C, E, G) Transcript age distributions are shown for each of the time points post-LPS, with the shift in density to the right (especially in the 2h, 4h and 6h conditions) being due to the prolonged higher transcription rate.

Plotting the transcript age distributions for several other top ranking genes for

increases in abundance 6 hours post-LPS - such as NIBAN1 and GNB4 (Figure 4.7D,F) - revealed a similar recording of past transcription (Figure 4.7E,G).

To test whether the differential gene ages were capable of capturing high-level regulatory changes, I first tested whether any pathways were significantly enriched with differentially aged genes with GO BP analysis. However, no pathways were identified for any of the contrasts, even when the $|\log_2(\text{AgeFC})| \geq 0.5$ requirement was removed entirely. As before with the HEK cells exposed to heat shock, this was predicted to be in part due to the small number of genes amenable to age analysis (mean $n = 774$ genes per contrast)

Inspired by the ability of hierarchical clustering in the HEK induction experiment to identify the stimulation applied (Figure 4.2B), clustering was also explored as a means to identify patterns of differentially ageing genes. Given that the steady state distribution of transcript ages for most genes is predicted to be exponential, and that perturbations to transcription rate or decay rate result in departures from this distribution, I reasoned that past changes in transcription may be detected as deviations from an exponential distribution - as observed for EIF2AK2, NIBAN1 and GNB4 in Figure 4.7C,E,G. For each gene, an exponential model of the form

$$N(\tau) = N_0 e^{-\delta\tau} \tag{4.1}$$

- where N is the number of transcripts of a given estimated age, τ, and δ is the decay rate - was fit to the transcript ages in the 0h time point and the decay parameter, extracted (as described in Chapter 3 section 3.5). This was used to generate a predicted distribution of transcript ages in the 6h time point, using the observed density of transcripts in the first age bin as N_0. These predicted values were then subtracted from the observed values to give the deviation from the expected underlying exponential distribution. Genes were filtered to those that had sufficient numbers of old transcripts for a potential change in past expression to be detected (a heuristic of ≥ 20 reads in the 6-8h age bin was used) - leaving 57 genes. K-

means clustering was used to produce 10 clusters of genes, which revealed patterns in transcript age distributions, such as clusters III and IV, which had fewer 2-4 h old transcripts than expected, and clusters VII, VIII which had more transcripts than expected in the 6-8h and 4-6h bins, respectively (Figure 4.8A). Trajectory plots of the mean ages of these genes over the course of the experiment identified some trends: such as cluster X being characterised by a decrease in the 2h condition relative to the 1h, followed by a gradual increase, and cluster VII which had relatively little change in mean age between the 2h and 4h post-LPS conditions followed by a sharp increase in the 6h - suggesting that the transcript age distribution encoded some of the transcriptional history. However, substantial intra-cluster variability remained and there was no clear correlation between the inferred transcriptional history from the heatmap and the changes in the mean ages over time. Indeed, the three genes shown in Figure 4.7B-F that were thought to have recorded a rapid increase in transcription 2 hours after LPS treatment were found in three separate clusters here. Future efforts may seek to improve the method for quantifying deviations from the expected distribution or to manually curate genes into groups that display the same change in expression over time and then testing whether clustering can rediscover the groupings.

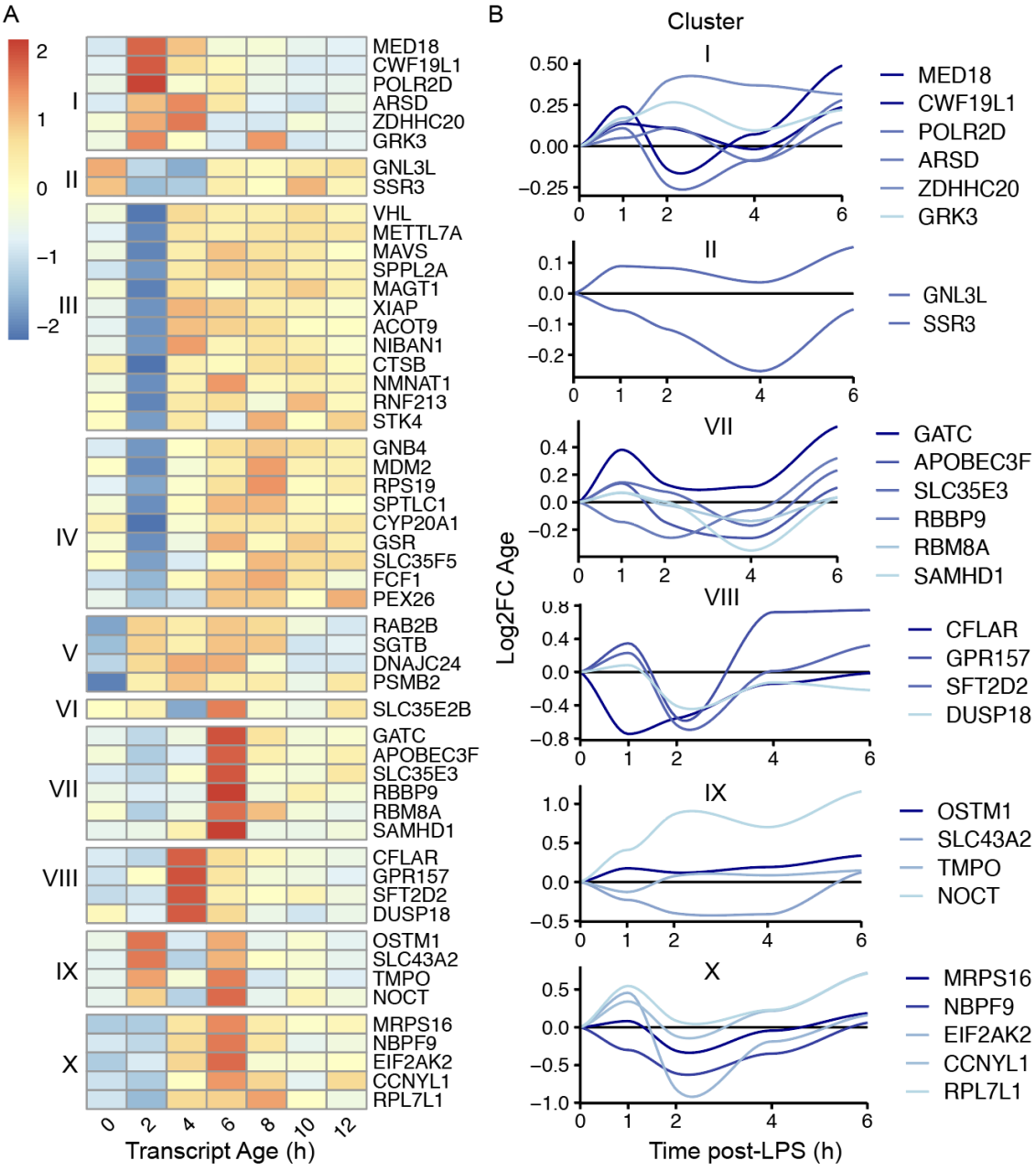

Figure 4.8: **K-means clustering reveals genes with shared changes in their transcript age distributions 6 hours post-LPS.** (A) For each gene, the mRNA decay rate before LPS addition is estimated by fitting an exponential decay model and then used to generate an expected distribution of transcript ages for the 6h condition. The expected values are then subtracted from the observed values to give the deviations from the exponential distribution, which are plotted as a heatmap and z-scaled by rows. Red values indicate large positive deviations, blue values indicate large negative deviations. The deviations are k-means clustered into 10 groups, which identify shared patterns. (B) For Clusters I, II, VII, VIII, IX and X, trajectory plots showing the $\log_2$(FC) in age relative to the 0h are shown for the genes in the cluster, identifying some shared historical changes in mean age.

In summary, LPS stimulation of monocytes served as an excellent testing ground

for ERA and demonstrated its ability to extract extra transcriptional histories from bulk long-read RNA-seq data.

4.5 ERA is a broadly applicable method that works across cell lines, primary immune cells, organoids and resected human brain

Having demonstrated the ages of pools of transcripts can be determined in unmodified monocytes, we sought to apply ERA in two other settings relevant for biomedical research: organoids and resected solid tissue, which we used intestinal organoids and organotypic human brain tissue.

Both samples were short-read sequenced with Illumina and ages were calculated using the same calibration dataset as before. The resected brain sample yielded very few aligned reads which consequently limited the number of editing sites detected and introduced a bias for more highly edited sites (since unedited sites and those with fewer than five reads do not pass preprocessing) (Figure 4.9A,D). Even still, 3,334 of the fit sites from the calibration dataset were detected in the resected human brain tissue and these were used to generate mean age estimates for 111 genes. Comparing the mean ages for these 111 genes with the ages for the same genes in the iPSC-derived cortical neurons revealed a strong correlation of the rank order (Spearman's Rho = 0.68), however the values correlated poorly (Pearson correlation coefficient = 0.26) - with the brain ages exhibiting older mean ages on average (Figure 4.9B). This result is consistent with the expected bias towards older ages that the low sequencing depth would introduce, as described above. Indeed, filtering the genes to those that have >5 sites produced a list of 38 genes which had strong correlation both in the rank order and the age values (Figure 4.9C). The same analysis of the human intestinal organoids (Figure 4.9D) produced 1,008 genes (Figure 4.9E) of which 551 had >5 sites and which displayed a stronger correlation between the

values (Pearson correlation improved from 0.47 to 0.60) (Figure 4.9F). However, unlike the ages from the brain, which skewed older than the iPSC-derived cortical neurons, the human intestinal organoids displayed younger on average ages.

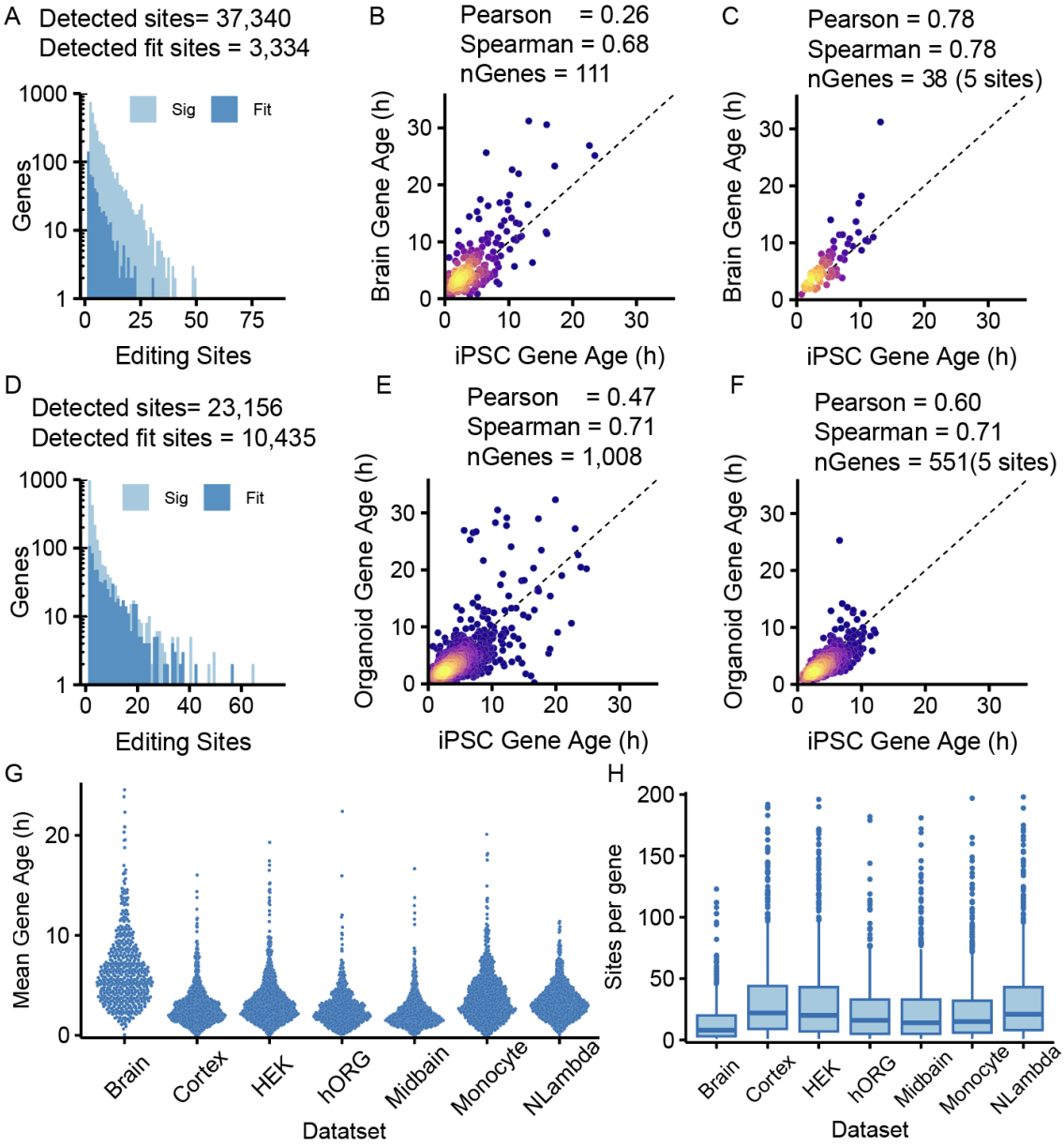

Figure 4.9: **ERA is widely applicable across a range of human tissues and cell lines.** (A) Resected human brain tissue was short-read sequenced and harboured thousands of the sites identified in the calibration data. Mean ages produced by ERA correlated well by rank with the iPSC-derived cortical neuron calibration data (B) and by value when subset to genes with at least 5 sites (C). (D, E, F) The same as A, B and C but for human intestinal organoids. (G) Summary plot of the gene ages calculated by ERA across the 7 major datasets in the project, shown as a beeswarm plot with each point denoting the mean age of a gene. (H) The number of fit sites per gene detected in each dataset, showing the low number retrieved from the resected human brain sample.

As shown in Figure 4.9G, which plots the mean gene ages for all of the main datasets from this thesis, some samples have higher mean ages than others and other differences to their gene age distributions. Some of these differences may be due to different levels of editing activity in different cell types (as was observed for the comparison of the NLambda HEK calibration dataset and the Cortical neuron dataset), however, without performing a calibration experiment for each cell type, the difference in mean age being due to different mRNA half-lives between samples cannot be determined. Indeed, the distribution of fit sites detected per gene across the different datasets (Figure 4.9H) did show a substantially lower number for the human brain sample which may be a factor in the higher ages observed - yet the monocytes, which have fewer sites per gene than many of the other datasets, displays the next highest ages after the brain samples.

4.6 ERA identifies the age of single RNAs in single cells

The popularity of single cell RNA-sequencing has grown in recent years due to its ability to resolve differences in gene expression between individual cells that are lost in the averaging process of bulk RNA-sequencing. My collaborator, Aaron Wagen, prepared 96 individual HEK cells for long-read sequencing by FACS sorting each cell into an individual well on a 96 well plate. The samples were sequenced on a PromethION24 sequencer across six R10.4.1 flow cells yielding approximately 44m long-reads per flow cell ($\sim$2m raw reads per single cell).

The editing of each read was extracted using Algorithm 1 as for all other long-read datasets. 9 cells were excluded due to having too little data (barcodes 24, 28, 18, 53, 15, 89, 62, 42 and 43 had <5,000 edited transcripts extracted). The MLEs of transcript ages were then calculated across all the data yielding an average of 6,835 MLEs per HEK cell (Figure 4.10A) across an average of 275.3 genes per cell (833 genes were found across the entire dataset) (Figure 4.10B). The number of MLEs

per cell was only moderately predictive of the number of genes per cell (Pearson's correlation coefficient = 0.423), suggesting that many genes had very few MLEs and were liable to drop out of the data between single cells. This was confirmed by visualising how many genes were shared by a given number of cells (Figure 4.10C) - showing that most genes were only shared by between 0 and 10 cells. In the analysis of the monocytes, genes were filtered to those that had 70 or more 'states' (that is - how many unique MLE values in the data) yet that caused severe drop out of the data here. Instead, a threshold of 10 or more states was used, which still reduced the number of genes substantially but provided some increased reliability of the mean ages (Figure 4.10D).

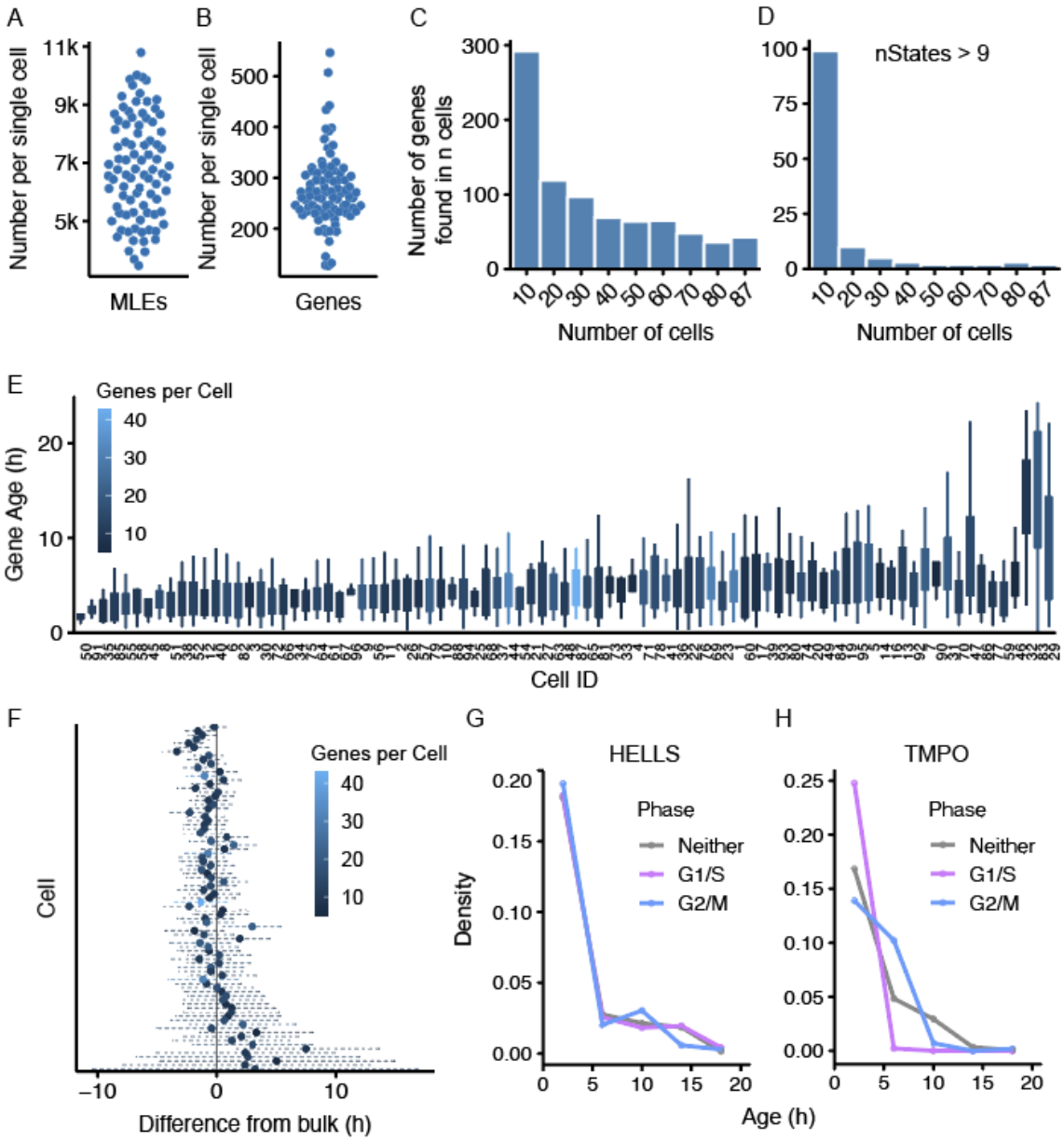

Figure 4.10: **ERA measures the ages of genes in single-cells.** (A) 96 HEK cells were long-read sequenced (ONT) and the ages of thousands of individual transcripts determined by MLE (A) across hundreds of genes (B). (C) Low read counts per gene produced sparse data in which few genes were shared by many cells. (D) Same as C but genes in each cell were filtered to those that had at least 9 unique MLE values. (E) The ages of the genes in D are shown as boxplots for each cell, ordered by mean age and coloured by genes per cell. Boxplots show the interquartile range (IQR), with whiskers extending to the maximum value within 1.5 times the IQR. Outliers are not shown and the y-axis is truncated to 25 h for visualisation purposes. (F) The mean age of each gene in each single cell was compared to the mean age of that gene in bulk sequencing of HEK cells. The points show the mean and the dashed lines extend to one standard deviation in either direction. The data is ordered by standard deviation with the lowest at the top, and coloured by the number of genes in each cell (row). (G & H) Distributions of transcript ages for two cell cycle associated genes in three categories of cells: G1/S, G2/M or neither (as described in Methods). (H) TMPO displays younger transcripts in single cells identified as being in G1/S phase and more older transcripts in cells associated with G2/M phase.

For each of the HEK cells, the mean ages of their genes were visualised as boxplots, from which two cells - barcodes 29 and 83 - were identified as anomalous, having a mean age of 38.1h compared with the rest of the cells (which had a mean age of 5.1h), and were thus removed from further analysis.

The mean ages of the single cells were compared with the mean ages from bulk long-read sequencing of HEK cells (specifically, the 0h and 4h control conditions from the heat shock experiment described in Section 4.3). The age of the gene in the bulk HEK was subtracted from each gene in each cell, producing a set of ΔAge values, which are shown in Figure 4.10F as mean and standard deviation. Whilst the spread of the age differences was large, the mean age values for each cell and over the 85 cells in aggregate were in extremely good agreement with the bulk, with a mean age of 5.13h across the single cells compared with 5.14h for the bulk.

Even though the single cells were not subject to a stimulation, they were actively dividing and thus some were expected to be at different stages of progress through the cell cycle. My collaborator Aaron ordered the cells by the ratio of the TPM values for two sets of genes that are considered to be G1/S associated and G2/M associated, respectively (Tirosh et al., 2016), and I performed analysis on the ages. Of the 98 genes that made up the G1/S phase and G2/M phase lists, only 6 were detected in the single cell long-read data and only HELLS (a DNA helicase) and TMPO (thymopoietin) were present in a sufficient number of cells at sufficient depth. To test if these two genes displayed differential transcript ages in cells more associated with G1/S phase or G2/M phase, the top 15 cells by rank of the expression of G1/S or G2/M associated genes were selected and the distributions of their transcript ages plotted (Figure 4.10G,H). Cells that weren't classed in either the top 15 G1/S or G2/M cells were simply labelled as 'Neither'. For HELLS, there was little observable difference between the three distributions (Figure 4.10G). However, for TMPO, transcripts from the top G1/S cells appeared to be younger than those from the G2/M cells, with cells labelled as 'Neither' being in the middle of the two (Figure 4.10H). TMPO is thought to be expressed early in the cell cycle and so although it

was labelled as being in the G2/M set by Tirosh et al. (2016), here it appears that expression starts in the G1/S phase and transcription is already decreasing by the G2/M phase. However, the low counts of TMPO transcripts and cells sequenced should serve as caveats to this conclusion. Nonetheless, it would be of interest to see if future experiments with more data reproduce this profile for TMPO and reveal similar patterns in other cell cycle genes.

4.7 Discussion

In this Chapter I demonstrated the use of the age methods developed in Chapter 3 to interrogate gene expression dynamics in increasingly complex systems. The validation of the calibration and age methods described in Chapters 2 & 3 was limited to either measuring steady state, or transcriptional halting by the addition of ActD. It was not known whether ERA would be able to identify increases and decreases of gene expression in actively transcribing cells.

The simplest system in which to test this was to place specific, well-edited genes under the control of an inducible promoter. This was done with ATG14, BVES and ACBD7 under the control of the Tet-On system, which increases transcription upon addition of the chemical doxycycline (Dox). Using this system, the upregulation of gene expression (ON), downregulation of gene expression (OFF) and an increase followed by a decrease (PULSE) were modelled (Figure 4.1). For ON, the transcript ages in the time series of samples taken after Dox addition showed an increase in young transcripts, corresponding to increased transcription and a concomitant decrease in the mean age of the transcripts. This behaviour recapitulates findings from Rodriques et al. (2020) Figure S1B and S5C, which showed an initial decrease in the mean edits per read immediately following the application of a stimulus. After reaching a minimum 1 hour after Dox addition, the mean age of the transcripts then increased over the remainder of the time points, with this being observed as the density of transcript ages shifting to the right in the top row of Figure 4.1B.

The distributions in the OFF conditions had little observable change in their shape over the course of the experiment but the mean age did increase in all three genes. Of note, there was no 'Dox_0h' condition which hampered comparisons between the OFF conditions. The PULSE conditions largely repeated what was observed in the calibration experiments which were also treated with ActD (e.g. Figure 3.4D). However, the mean age changes between the conditions post-ActD were smaller than those observed in the calibration data, potentially due to the competing effect of doxycycline.

The plots in Figure 4.1B provide some insight into how the shape of the transcript age distributions is influenced by a stimulation. First, the faster the turnover of a gene's mRNA, the faster the shape of the age distribution will return to a steady state - as can be seen by the mean age of BVES (which had a mean age in the 0h condition of 1.4h) having already surpassed the 0h level by the 4h time point, whilst for ACBD7 (which had a mean age in the 0h condition of 3.8h) this level is reached sometime between the 4h and 8h time points. A similar observation was made by Rummel et al. (2023) in the context of metabolic labelling data. Second, interpretation of differential expression analysis typically assumes that an increase in counts corresponds to a gene being upregulated in one condition relative to another - and that the larger the increase in counts, the larger the upregulation. The same is not true for a change in age between two conditions. For instance, samples separated by 8 hours following each of ON, OFF and PULSE stimulations all display an increase in mean age relative to the starting condition. Repeating the clustering in Figure 4.2B but using the 8h (ON), Dox_8h (OFF) and Act_8h (PULSE) conditions showed that the age distributions clustered by gene rather than by stimulation for the ON and OFF conditions (the PULSE was trivial to distinguish and clustered separately). However, for each of the genes the OFF condition had an older mean age, enabling distinction from the ON condition. As discussed above, the OFF stimulation was weak compared with the change observed when transcription is halted entirely by ActD, and so a stronger OFF stimulation may have produced

clearer results.

To move beyond an induction system to measuring expression changes from endogenous regulation, we sought a stimulation that would cause a large change to as many genes as possible in HEK cells - the system we had used to characterise most of the method thus far. Heat shock was selected, with the experiment carried out by my collaborator Ali. In general, very few of the genes that were differentially expressed (as determined by DESeq2 (Love et al., 2014), a common software used) were also well edited. Performing differential age analysis (defined in Methods) identified a greater number of genes - although these were, by definition, restricted to well-edited genes. As shown in Figure 4.4B,C,D,E, genes that were differential by age analysis but not DESeq2 displayed interesting changes to the ages of their transcripts, with CFLAR identified as potentially being stabilised (Figure 4.4D, bottom row).

Since ERA works in unmodified human cells, in principle it should be applicable to primary cells. Stimulating primary human monocytes with lipopolysaccharide (LPS) induced a substantially stronger transcriptional response than the heat shock of the HEK cells. This enabled the detection of transiently upregulated genes from the 6h post-LPS time point, such as the ribosomal proteins RPS19 and RPL37A (Figure 4.6. It also identified transcripts of CYP20A1 - the only cytochrome p450 protein for which the function is unknown - as being destabilised, implicating it in the response to LPS.

Many well-edited genes in the monocytes had much larger increases in expression than measured in either the Tet-On induction experiment or the HEK heat-shock. In several genes, this signal was strong enough that it could be detected in the transcript ages from a sample taken several hours after the initial increase (Figure 4.7C,E,G). This provides a proof of concept that the transcriptional recording described by Rodriques et al. (2020) is also possible for endogenous genes using only endogenous editing. As seen in Figure 4.8 , in the final time point (6 hours post-LPS), clustering the genes by the deviation of the transcript ages from the expected distribution

reveals common behaviour. Furthermore, plotting the trace of the mean ages of those genes at the intermediate time points reveals correlations within many of the clusters, again suggesting a degree of transcriptional recording. However, there was heterogeneity in the performance across the clusters and it was difficult to find a correspondence between the patterns in the heatmap of Figure 4.8A and the trace in 4.8B.

Despite the high drop off rates from raw sequencing reads to MLEs, whether ERA would be applicable to single cell RNA-sequencing data was investigated. 96 individual HEK cells were sequenced from plates on a PromethION24 machine, yielding 87 cells with sufficient reads. As expected, the representation of well-edited genes across the single cells was sparse and few genes were identified in all 87 cells (Figure 4.10C,D). For 85 of the cells, the mean ages of the genes agreed extremely well with the mean age of the same genes in HEK cells that underwent bulk long-read sequencing, although the variance was substantial (Figure 4.10F). Interestingly, 2 of the cells (barcodes 29 and 83) exhibited a ten-fold higher mean gene age than the other 85 cells. These two cells had relatively few MLEs (barcode 29 ranked 6^{th} last, barcode 83 18^{th} last) and also had the 2^{nd} and 3^{rd} lowest numbers of genes. However, barcode 32, which had the fewest genes and 4^{th} fewest MLEs had a mean gene age of 8.9h, which was much more similar to the other cells than barcodes 29 and 83. Inspecting the bam files for barcodes 29 and 83 using the Intergrative Genomics Viewer (IGV, Robinson et al. (2011)), showed noticeably higher A-G mismatches in several well edited genes, suggesting that either the editing rates in these cells was markedly higher than in the other 85 or that something occurred during the library preparation which resulted in a high proportion of highly edited transcripts in the final cDNA library. We did not use unique molecular identifiers (UMIs) so the level of duplicated reads was not assessed. Although the activity of ADAR1 can be increased - most typically by the upregulation of the inducible p150 isoform - the observed increase in editing is greater than has been reported elsewhere (Chung et al., 2018) and since these are HEK cells at rest, no upregulation of p150 would

be expected at all. Thus, the cause of the abnormally high ages in barcodes 29 and 83 remains a mystery.

Chapter 5

Discussion

5.1 Contextualisation of major findings

The research presented in this thesis sought to answer two overarching questions: first, were the rates of endogenous A-to-I editing in human cells sufficient to calculate the age of mRNA transcripts and second; could the ages of these transcripts be used to infer changes to gene expression.

5.1.1 Characterisation of editing rates at endogenous loci

ADAR mediated A-to-I editing is a large field of research, yet most studies simply report the number of editing sites in a given sample or the observed level of editing at the sites. These measurements are related, since whether a particular locus is classified as an editing site, a single nucleotide variant (SNV) or a sequencing error is often determined by whether the number of A/G mismatches is above a certain threshold (1-5% of mapped reads is typically used as a minimum threshold). Thus, the number of editing sites detected depends on the threshold set for the minimum level of editing. However, the observed snapshot *level* of editing at a site does not necessarily reflect the underlying editing *rate* at that position. This distinction has been largely neglected, and no studies have sought to determine the underlying editing rates (that is - how quickly ADAR edits at a given position) in a high-throughput way. Here, measuring the underlying editing rates was a necessary step in seeking to adapt the method of Rodriques et al. (2020) to the endogenous setting. To do this, the calibration methodology put forward by Rodriques et al. (2020) was

adapted to the endogenous setting to measure the editing rates of all expressed editing sites in parallel. This was done by halting the production of new transcripts with Actinomycin D, and then taking samples at pre-determined time intervals and sequencing. This enabled us to move beyond the snapshot measurement of editing *levels* to an understanding of the editing *rate* kinetics.

The hundreds of thousands to millions (depending on the cell line and sequencing depth (Table 2.1)) of sites identified in the steady state level (i.e. the 0h time point, just before ActD addition) could be classified by the rate at which edits accumulated at those sites. In Chapter 2, I first define a set of significantly edited sites ('Sig sites') which are those that significantly increase in editing level 8 hours post-ActD addition. These are then subset further to those that are well fit by the calibration equation (Equation 2.5), yielding the set of 'fit sites'. There were on the order of 30,000 to 160,000 fit sites in the transcriptome, depending on the cell type and sequencing depth used. Literature values for the number of editing sites detected human tissues range from 10^5 and 10^6 (such as in the REDIportal database of human ADAR edits, Picardi et al. (2017)), which is more similar to the counts of 'significant sites' here than to the counts of 'fit sites'. This is expected, since the methodology for identifying 'significant sites' is more similar to literature methods than the method for 'fit sites'.

Bazak et al. (2014) showed that as the sequencing depth at a putative editing locus increased, the rates of both type I and type II errors for labelling the locus as an editing site decreased - in agreement with the findings here that higher sequencing depth identified more sites. I found that increasing the number of reads preferentially discovered sites in introns, which was thought to be caused by the general lower sequencing coverage of these regions due to most sequenced transcripts already having had their introns spliced out (Mortazavi et al., 2008). Comparing the editing rates of sites between calibration datasets showed a strong correlation of the editing rates, implying that the editing rate of a site is robust, largely intrinsic and probably derives from the surrounding RNA sequence. Although editing rates have

not been determined before, it is well known that ADAR1 editing sites are found in *Alu* elements and in dsRNA. Although interesting, investigating the sequence determinants of editing rates was not a focus of this work. However, if it is possible to predict editing rates from sequence alone, then this would remove the need for further calibration experiments and enable age analysis on genes that were not expressed during the calibration experiment. Since many regulatory genes are only expressed in response to specific stimuli, at present they cannot be studied using ERA.

Although others have reported that cell lines such as HEK293 display far less editing than other human cell types, I did not find large differences between the editing *rates* at shared sites between in iPSC-derived cortical neurons and HEK cells (Schaffer et al., 2020), although the total number of sites was indeed higher in the iPSC neurons. In summary, the results presented here constitute the first transcriptome-wide measurement of endogenous editing rate kinetics, which are robust and correlate highly between cell types.

5.1.2 Endogenous A-to-I editing encodes the passage of time

Having determined that these rates were robust and that there were tens of thousands of sites with an expected editing time between 0.5h and 100h (corresponding to the editing rate, λ, thresholds of $0.01 \leq \lambda < 2$), I then demonstrated that the editing rates could be used to infer the ages of transcripts and populations of transcripts. Two models of Endogenous RNA Age (ERA) were developed: per-transcript (which has some conceptual similarity to the method of Rodriques et al. (2020) yet is distinct in being a likelihood-based approach) and per-site, which was entirely novel. Per-transcript ERA estimates transcript age from individual long-reads, whilst per-site age estimates the mean age of the population of transcripts mapping to a given site. Using either of the two models, the mean age of the transcripts of a gene can be calculated and these accurately recovered the known duration of time elapsed between sequencing steps in the calibration experiments. Of note, the age estimates

from ERA are in true-time, in contrast to the pseudo-time estimates from RNA velocity methods (La Manno et al., 2018; Bergen et al., 2020). This demonstrated that endogenous A-to-I editing rates were sufficient for accurate transcript age estimation - answering the first of the two major research questions.

5.1.3 ERA estimates mRNA synthesis and decay kinetics

ERA also enabled mRNA half-lives to be calculated simply by multiplying the mean age by ln(2) or by estimating the decay rate from the distribution of transcript age estimates using

$$N(\tau) = N_0 e^{-\delta\tau} \tag{5.1}$$

(see Section 3.5 for details). The two most popular methods for estimating degradation rates (as introduced in Chapter 1) - transcriptional arrest and metabolic labelling - both require dedicated experimental protocols, typically involving multiple time points (although metabolic labelling can estimate half-lives from a single condition (Rummel et al., 2023)). Therefore, to the best of my knowledge, ERA is the first method that can measure half-lives in absolute time from standard RNA-seq data. Whilst deep long-read sequencing is required for per-transcript ERA (which is an uncommon sequencing choice), per-site ERA can be applied to short-read sequencing and will generate half-lives for genes at steady state mRNA expression (the steady state assumption produces the exponential relationship between transcript counts and age, which is required in order for $t_{1/2} = \ln(2) \times$ (mean age) to be true).

For genes where the transcript age distribution is well fit by Equation 5.1, it is possible to estimate not only the half-life (from δ) but also the synthesis rate from N_0 (in practice the normalised counts of transcripts in the first age bin are used). However, since the synthesis rates have relative units per hour, they are only informative when compared between conditions and not when compared between genes. This is similar to how synthesis rates are analysed from metabolic labelling data such

as with GrandR (Rummel et al., 2023), where the fold change of synthesis rates is analysed.

5.1.4 ERA characterises gene expression changes to stimulus

To see whether ERA could identify changes in gene expression (the second major research question), three well-edited genes were placed under inducible control and used to mimic upregulation and downregulation of gene expression by a promoter. Induction (ON) of expression by Doxycycline (Dox) was detected as an initial decrease in mean age due to new transcription followed by an increase in mean age as more time passed since Dox was added. This behaviour was also observed by Rodriques et al. (2020) for the mean number of edits per RNA timestamp, which first decreased with addition of Dox and then began increasing shortly after ActD addition (shown in Supplementary Figure 1B of their work).

In our experiment, for samples where Dox was then removed (simulating downregulation, OFF) the mean age increased as time passed. The OFF stimulation appeared weak, and a stronger OFF stimulation may have been achieved had the Tet-Off system been used specifically for this condition, rather than relying on the deactivation of Tet-On. Whilst the inductions produced the expected behaviours, it was clear that the increases and decreases in mean age alone were not necessarily sufficient to distinguish between the increased or decreased expression of a gene. This is in contrast to differential expression analysis, where the change in normalised counts is directly indicative of increased or decreased transcript abundance (Love et al., 2014). It was therefore reasoned that combining age information and abundance information would help to distinguish different changes to gene expression.

This hypothesis was tested in HEK cells exposed to heat shock, which caused differential expression of hundreds of genes. To integrate the transcript abundance information with the transcript age information, rather than generating a distribution of the *density* of transcript ages, distributions of the *normalised counts* of

transcripts by age were created instead (Methods, Section 6.3.2). This enabled statements about the relative change of transcripts of specific ages between conditions. For example, MFSD11 was identified as having a decreased transcription rate 4 hours after heat shock, since the proportion of the youngest transcripts was lower than before the heat shock (Chapter 5, Figure 4.4D).

This integrated metric of 'normalised counts by age' had the benefit of enabling comparison of transcription rate changes between conditions whilst still being able to approximate half-lives, since the mean age value was unaffected by the normalisation of counts. Existing methods to obtain similar information on transcription and degradation rate changes - such as pulse-chase metabolic labelling - involve complex experimental and bioinformatic workflows (Rabani et al., 2011; Rummel et al., 2023). ERA, by contrast, requires no special experimental procedures (just long-read sequencing) but is limited to well-edited genes, whilst metabolic labelling methods are more broadly applicable. The methods proposed by Jürges et al. (2018) and Rummel et al. (2023) have sophisticated models for handling error when estimating the kinetic parameters, which are underdeveloped for ERA. However, due to the conceptual similarities between metabolic labelling and A-to-I editing, it may be possible to adapt their methods to ERA.

Although promising, the HEK heat-shock experiment was limited by the small number of differentially expressed well-edited genes. To study a larger transcriptional response and to demonstrate ERA's applicability to primary human cells, a new experiment was performed in which monocytes were stimulated with lipopolysaccharide - inducing broad transcriptional reprogramming (Chapter 5, Section 4.4). In this setting, a range of temporal gene expression changes were suggested by ERA, including transient upregulation (RPS19, RPL37A), transcript destabilisation events (CYP20A1), changes to transcription rate (VHL) and the ability to record past events in the distribution of transcript ages (EIF2AK2, NIBAN1, GNB4). Of note, EIF2AK2 (also known as PKR) is one of critical downstream targets of cytosolic ADAR1-p150, and it has been shown that ADAR1 binding to dsRNA inhibits PKR

activation (de Reuver et al., 2022).

Chan et al. (2020) suggested that editing of EIF2AK2 transcripts influences its mRNA abundance levels, and that this may be mediated by the RNA binding protein ILF3 - which binds to EIF2AK2 transcripts. Indeed, the 3'UTR of EIF2AK2 was identified as harbouring a large number of fit sites in the calibration experiments. However, given that apparent changes in editing can in fact be caused by changes to abundance (for instance, mRNA stabilisation events increase abundance and produce an apparent increase in editing levels, without the need for a change in underlying editing rate) the conclusions of Chan et al. (2020) should be taken with caution. It may be the case that the editing level is serving as a proxy measurement of a change to another process (such as decay rate) and thus the change in editing is 'effect' rather than 'cause'.

This leads into a general observation that it is important, not only in the work here but also for the literature, to distinguish between editing levels and editing rates. If future experiments and analysis corroborate that editing *rates* are stable between cell types (as was suggested in Chapter 2) and yet changes to editing *levels* are still observed, then further analysis may reveal these to be the result of degradation changes. It is known that increased expression of ADAR1p150 in response to interferon results in increased observed editing levels (Chung et al., 2018) and this may be due to changes in the underlying editing rates. However, if all the editing rates are scaled by the same factor then: 1. this can be accounted for and 2. this does not explain how some sites could display increased editing whilst others display decreased - suggesting that these events are caused by a different mechanism. Therefore, the paradigm of editing as an underlying process rather than a static level provides a critical lens through which to view results that implicate changes in editing levels as a cause rather than effect.

In conclusion, whilst many of the more complex changes observed in transcripts ages were only qualitatively assessed, the results provide a proof of concept for transcript

ages being able to decode endogenous transcriptional histories: thereby answering the second major research question.

5.2 Strengths, Limitations and Future Directions

5.2.1 The Calibration Process

The aim of the calibration process in Chapter 2 is to determine a list of editing sites which can be used in ERA. The first challenge is identifying editing sites. This is an unsolved problem and methods are frequently published to determine whether observed mismatches in RNA-seq data correspond to real editing sites (Piechotta et al., 2017, 2021). A similar problem also exists for metabolic labelling methods, where the low incorporation rates of 4sU make the results sensitive to false positives Jürges et al. (2018). Here I used JACUSA2 to identify editing sites which, as discussed in Chapter 3 Section 3.4.3, filters sites to those that have at least five reads and at least one edit. Whilst this is useful for the task of determining the statistical significance of an editing event, it results in the loss of valid 'non editing' events at loci that have already been characterised as editing sites. This issue could be easily resolved by using a lower-level function to retrieve the editing information from the .bam files, such as the SAMtools mpileup function.

After determining which sites are *bone fide* editing sites, the next issue is determining which should be used for age inference. Here, the calibration equation (Equation 2.5) is fit and sites are filtered based on a combination of R^2, λ, observed editing level and the number of calibration time points the site is detected in. Future work may make substantial improvements to the rigour of this filtering step. Furthermore, since there is substantial drop off from 'sig sites' to 'fit sites', there is an open question of why some sites are *not* fit by the calibration experiment. There are several obvious answers: low read counts, genomic SNVs or the site being spliced early and therefore not available for editing at later time points. However, further analyses of these sites may discover one or more distinct categories into which they fall.

One hypothesis of interest is whether some sites are only edited in the nucleus: either because they are exclusively edited by ADAR1-p110 (and thus very unlikely to be edited after nuclear export) - or because they reside in introns that are later spliced and the editing is occurring co-transcriptionally (again, likely a p110 editing site). Future work may determine if fit sites tend to be targets of p150 or p110, and indeed integrating information on spliced and unspliced transcripts (especially from the long-read data) may provide clues (for instance, it would be unusual to see edits at a slow editing site on a fast spliced intron). A recent study that identified sites as being targets of either p150 or p110 found that most editing sites (62%) were edited solely by p150 and that the rest were edited by either isoform (38%) (Sun et al., 2021). They did not identify any sites that were the target of p110 exclusively. This suggests a binding model where the p150 isoform can bind any p110 target, whilst having a further set of p150-only targets due to its Zα domain (Sun et al., 2021). Chapter 1 Section 1.2 detailed the amount of time that transcripts spend in each stage of their lifecycle: during transcription, in the nucleus prior to export and then in the cytosol. These values have implications for editing, especially when it is viewed as a probabilistic process over time as it is in this thesis. The longer a transcript spends in the nucleus, the more time there is for it to be edited by p110. Similarly, the more time spent in the nucleus, the more time for p150 to edit. This is corroborated by the results of Sun et al. (2021) where p150-only sites are enriched for 3'UTR sites and depleted in intronic sites vs p110/p150-shared sites. The authors suggest that the presence of p150-only sites enables the cell to not just increase the editing levels in interferon response, but to also increase the number of sites in total. When the effects of filtering threshold and the probabilistic nature of editing are taken into account, it seems likely that the 'new sites' were in fact always targets of ADAR1, but only pass the filtering criteria once the underlying editing rates are above a certain level. To test this theory, future work should overexpress p150 and see whether the editing rates increase gradually or whether some sites 'pop in' to the data. This would also have implications for the claim of Bazak et al. (2014) that there are over

100 million editing sites across human genes - far more than the number identified in the calibration experiments of Chapter 2. A recent study from Deng et al. (2025) on the structural and sequence basis of ADAR1 editing suggest that ADAR1 has moderate specificity for the length of dsRNA that it edits and for the sequence surrounding the editing sites. Nonetheless, the sheer quantity of *Alu*-elements in the human genome would still enable the quantities of editing proposed by Bazak et al. (2014). Future work then, should focus on identifying the mechanisms through which editing levels increase in response to interferon. In particular, whether it is simply that the concentration of ADAR1 is limiting under constitutive expression, whether there is a co-operative binding mechanism for multiple ADAR1 proteins binding the same dsRNA (it is already known that ADAR1 forms homodimers on dsRNA (Mboukou et al., 2024)) or whether there is a cofactor or some other regulatory mechanism that increases the editing activity of ADAR1. In conclusion, future research of the binding and editing mechanisms of ADAR1 may benefit substantially from measuring changes in the underlying editing rate.

5.2.2 Per-site and per-transcript age methods

As was discovered in Chapter Section 3.4.1 of 3, the number of 'fit sites' in the calibration data are far in excess of the number of sites that are typically found in subsequent experiments. This poses a serious dilemma of which fit sites to use. I used a heuristic of only 3'UTR sites in genes with at least five 3'UTR sites, which was motivated by most of these sites likely being in *Alu*-repeats and thus sharing a biological mechanism for why they are edited. However, this limited the number of well-edited genes and did not completely remove poorly performing sites (as evidenced by the fraction of zero-edit reads present 16 hours post-ActD). Defining more rigorous criteria for what constitutes a 'well-fit site' should be a priority for future work. This may include greater use of sequence context: for instance, whether a site is in an *Alu*-repeat or is only found in a particular isoform.

In general, considering isoforms should be an area of focus for future work. Here,

all data was aligned against the reference genome and no attempts were made at assigning reads specifically to specific isoforms. For the calibration process, long-read sequencing and alignment to the transcriptome may identify editing sites that have different editing rates on different isoforms. A long-read calibration dataset could also be used to test the validity of the assumption of independence of editing between sites, a key assumption of the per-transcript likelihood model.

Measuring the uncertainty of the per-transcript age estimates should be a focus of future work. In the current maximum likelihood estimation (MLE) implementation, I report a single value corresponding to the most likely age of a given read. In the first instance, a confidence interval (CI) should be calculated for the age parameter, which could be achieved by the using the likelihood ratio test which essentially calculates the CI from the curvature of the log-likelihood function (the stepper the curvature the smaller the CI).

5.2.3 Towards quantification of changes to mRNA kinetics

With a better understanding of the uncertainty in the transcript age estimates, future work should seek to improve the analysis methods for the transcript age distributions. Using the analogy of the metabolic labelling, ERA currently provides the proportions of young and old reads (although of course ERA estimates *how* old each read is). That serves as a crucial substrate for downstream estimation of kinetics, which here was quantitative for half-life, but qualitative for synthesis rate and for recording of past transcription events. A few specific avenues for development are suggested below:

In the work here, the transcript ages were binned and then these histograms compared between conditions. Future work should seek to replace the binning procedure - which introduces sensitivity to the bin width chosen - and to develop a set of statistical tests for hypothesis testing of changes to the age distributions. This would be critically important for any attempts to extract transcriptional history, as was

attempted in Figures 4.7 and 4.8. The Mann-Whitney U test implemented for mean age does so using the non-binned per-transcript MLE values, so is not sensitive to the binning procedure. However, it is sensitive to tied values and since the distribution of transcript ages is exponential (with most reads having an MLE of 0 hours), this may introduce a large tie-correction which leads to a conservative p-value. The effect of this on the analysis of differential age should be assessed in future work. In particular, since the occurrence of tied values is likely to be high (since there is a finite set of MLE values that a transcript with any given number of editing sites can produce), further changes to the statistical method should be explored.

Future work may also combine per-site and per-transcript information into a single age estimate. In Chapter 3 Section 3.4.4 it was observed that the estimates of the mean age of the population of transcripts from per-site age were more accurate than those from per-transcript age, especially at older mean ages (Figure 3.6A,B and Figure 3.7B). Thus, it may be possible to construct a Bayesian framework for transcript ages in which the posterior estimate of the age of each transcript incorporates both the MLE and the population mean generated from per-site ERA. The application of Bayesian frameworks to metabolic labelling has seen recent interest (Rummel et al., 2023) and it may be possible to transfer methods easily to ERA data.

With whichever distribution of transcript ages is chosen, future efforts should build on the work here on half-lives to develop methods for comparing log(fold changes) of transcript synthesis rates between conditions. The method chosen will depend on how the transcript ages are represented. Several ideas were explored during the course of this research, including comparing the N_0 values from the fit of the exponential decay model and comparing the normalised counts of unedited reads. However, ultimately neither were fully developed but may serve as starting points for future work.

5.2.4 ERA compared to metabolic labelling

The primary strength of ERA is that it works in unmodified cells using endogenous processes and therefore requires no experimental protocol. In this regard it offers advantages over metabolic labelling techniques such as BRIC-seq (Imamachi et al., 2014), SLAM-seq (Herzog et al., 2017) and Timelapse-Seq (Schofield et al., 2018), which require the addition of a labelling agent such as 4-thiouridine (4sU) and transcriptional halting methods for half-life measurement (Viegas et al., 2023). These methods do, however, work across a far greater number of genes (typically all expressed genes if sequenced at sufficient depth) and the metabolic labelling techniques are minimally disruptive to the cells, provided the labelling period is short (less than 24 hours) and the concentration of 4sU moderate (Rummel et al., 2023). ERA, whilst non-disruptive to the cell, has the potentially opposite issue of being disrupted *by* the cell under certain conditions - particularly in response to interferon which can increase the editing activity of the p150 isoform of ADAR1 (Chung et al., 2018). In this regard, ERA is similar to RNA velocity methods, which also utilise an endogenous process (splicing) to infer expression dynamics (La Manno et al., 2018; Bergen et al., 2020; Zheng et al., 2023). ERA would therefore be sensitive to changes in the editing activity of ADAR1. Depending on how this change manifests - whether all editing rates are simply scaled by some factor, or whether they are site specific - it may be possible to account for it. If editing rates scale globally, then the scaling factor can be determined and the editing rates used by ERA adjusted accordingly. However, the observed editing level of a gene expressed at steady state is a product of not only the editing activity but also the degradation rate. Therefore, a method would need to be devised to identify genes that have no change to their degradation rate and can be used to calculate the difference in editing rate. It may be possible to integrate multiple measurements from the same data RNA-seq data - total abundance, splicing ratios and editing levels - to investigate whether global changes in observed editing are the result of changed synthesis, decay or editing rates.

5.2.5 Increasing editing rates with engineered editors

Finally, the biggest limitation of ERA is the number of editing sites, and therefore the number of genes, to which it can be applied. Although the ADAR2(E448Q)-NLambda construct was abandoned due to not appreciably increasing useful editing activity, using engineered editors remains the most promising avenue to increase the number of genes that ERA can measure and the resolution of the method. However, moving away from endogenous editing would change the applicability of the method. Although the early results were reported here, an initial test with tRNA Adenosine Deaminase (tadA, a t-RNA specific adenosine deaminase that has shown utility in genetic editing work (Wolf et al., 2002; Gaudelli et al., 2017)) revealed a substantial increase in editing sites. Future engineering of this enzyme or others may yield a construct capable of editing most genes when expressed. However, such an editor may 'suffer from success' in that such a high degree of editing (especially in exons) may cause enough toxicity to render it unusable. Therefore, there may exist a theoretical upper limit to the resolution of age models using mRNA editing.

Whether the rates of other mRNA modification processes, such as m^6A methylation, exhibit similar properties as A-to-I editing was not investigated but one might imagine creating a unified model of transcript age given the relative modifications of a given mRNA. Given the rapid developments in direct RNA-sequencing with nanopores, capturing all of this information from a single RNA-seq run may soon be possible at scale.

5.3 Conclusion

The research presented in this thesis established that the principle of 'editing encoding time' extends to endogenous A-to-I editing in human cells. By modelling the rate at which those edits happen, it is possible to infer the ages of individual transcripts from RNA-sequencing data and to use those ages to study gene expression changes. Whilst this research project began with a strong method development

framing, many insights into the biology of A-to-I editing and the kinetics of mRNA turnover were produced when applying the method. The rich and complex life of a human mRNA transcript is such that I conclude this thesis with more questions than I started with - but also with a greater understanding and new tools of analysis at my disposal. Through the avenues of future work I have detailed in this Chapter, and due to the activity in the field in which the research sits, I believe that the work presented in this thesis is an important contribution towards our understanding of RNA editing processes and the dynamics of gene expression.

Chapter 6

Materials and Methods

This Chapter details the methods that I used to conduct the research presented in the thesis. Since my research was undertaken with collaborators, who performed all of the 'wet lab' experiments, I have included details of their methods (as written up by us in the preparation of a manuscript) in the Appendix that follows this Chapter, if of interest. Unless otherwise specified, all analysis was undertaken by me in R 4.1.0 (https://www.r-project.org/).

6.1 Processing of RNA sequencing data

6.1.1 Short-Read Data Processing

Bcl files resulting from short-read sequencing (Illumina technologies) were converted to FASTQ files using bcl2fastq2 v2.20.0 (https://emea.support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html). The resulting FASTQ files were fed into a custom Nextflow (version 22.04.0) pipeline ("Twits", version GY220823) written by George Young, the steps are detailed below.

Reads were trimmed using cutadapt v3.5 (https://github.com/marcelm/cutadapt/) and aligned to the GRCh38.100 human reference genome using HISAT2 v2.1.0 (http:// daehwankimlab.github.io/hisat2/). The counts per gene were then quantified using salmon v1.4.0 (https://combine-lab.github.io/salmon/). Editing was quantified using JACUSA2 v2.0.4 (https://github.com/dieterich-lab/ JACUSA2) with `call-1 (detect)` settings with flags `-p 16 -P RF-FIRSTSTRAND -a I,S,Y -F 3844`. The resulting JACUSA files were converted to BED files using a short

custom Python3 script written by George Young.

6.1.2 Long-Read Data Processing

POD5 files from ONT sequencing were basecalled in Super Accuracy Mode (SUP) using `guppy_basecaller` from guppy v6.4.6-CUDA-11.7.0 on NVIDIA A100 and V100 graphic processing units. The resulting FASTQ files were demultiplexed using `guppy_barcoder` from the same version. PacBio data was received from Oxford Genomics Ltd. as processed FASTQ files.

Demultiplexed FASTQ files from ONT or PacBio were subsequently processed using the same pipeline, which was also implemented as Nextflow (v 22.04.0) pipeline by George Young. Reads were aligned to the GRCh38.100 human reference genome using minimap2 (https://github.com/lh3/minimap2) and quantified using IsoQuant v3.0.0 (https://github.com/ablab/IsoQuant). Edits were quantified as described for short-read data using JACUSA2.

6.2 Modelling of endogenous editing rates

6.2.1 Editing Rate Determination from Calibration Data

Experimental details of the calibration experiments can be found in Appendix 6.4.9.

'Calibration experiments' involve the halting of transcription by the addition of Actinomycin D (ActD) and the lysing and sequencing of samples at defined time points after ActD addition. ActD is added at $t = 0$ hours, and time points are referenced as being relative to this time, with the exception of the $t = 0$ hour time point itself, which was in fact sequenced before ActD addition. At each time point, the editing level at each site is defined as the fraction of edited reads divided by the total reads that map to that site. In practice the editing levels are quantified using JACUSA `call-1` as defined above.

Sites with significant increases in editing level 8 hours post-ActD addition were

identified using JACUSA2 `call-2`, which determines whether a site is significantly differentially edited between two conditions. The significance is reported as a Z score, where a Z score with magnitude greater than 1.96 is considered to be approximately equivalent to a p value of less than 0.05. Sites were filtered to include only sites for which the Z score greater than 1.96. This gives the list of significant sites - or 'Sig sites' for short.

To obtain the list of 'fit sites', first a linear model of the form

$$y_j = \alpha + \beta\, t_j + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2) \tag{6.1}$$

was fit to each site using the `lm` function from the R `stats v4.1.2` package, where y_j is the observed editing level at the j^{th} time point, t_j. ε is a noise parameter used internally by the function. Any sites with a non-positive gradient, $\hat{\beta}$, were removed since these were assumed to have no accumulation of edits over time.

The full theoretical framework for modelling the accumulation of edits over time can be found in Chapter 1 Section 2.2. The key equations for methods are

$$y_i = 1 - e^{-\lambda_i t} \tag{6.2}$$

and

$$y_i = 1 - a_i e^{-\lambda_i t}, \tag{6.3}$$

where y_i is the observed editing level at site i at time t. λ_i is the editing rate at the site and a_i is the fraction of edited reads at site i at time $t = 0$.

Equation 6.3 is fit, for all time points from the calibration experiment that are less than or equal to 12 hours post-ActD addition, to each site, i, that passed the linear model filtering above, with a_i fit as a free parameter using the `nls` function from the R package `stats` v4.1.2, using starting values $\lambda_i = 0.01$ and $a_i = 1 - y_i(t = 0)$.

The results of the `nls` fit were filtered based on the following criteria to generate the list of 'fit sites':

- $R^2 \geq 0.4$
- $2 \geq \lambda \geq 0.01$ h^{-1}
- Each site must appear in at least four time points from the calibration dataset.
- The predicted age of the site at the $t = 0$ calibration time point (i.e., prior to ActD addition) is less than 25 hours.

Sites were annotated as belonging to a specific gene if their genomic coordinates mapped between the start of the 5'UTR and the end of the 3'UTR and were on the same strand. Sites located in regions annotated with more than one gene on the same strand could not be assigned unambiguously and were removed from analysis. Sites mapping to intergenic regions were also removed from the list of fit sites.

In later analysis, the list of 'fit sites' was further filtered to contain only sites annotated to the 3'UTR for genes that had at least five sites in the 3'UTR.

The calibration datasets from hiPSC-derived cortical neurons and HEK293 cells (transfected with ADAR2(E488Q)-Nlambda) were combined to increase the set of genes that could be analysed in downstream analysis. Since the editing rates in the cortical neurons appeared globally lower than those in the NLambda-HEK calibration, the datasets were combined by calculating the mean ln fold change at all shared sites, scaling the editing rates at the hiPSC-only sites by this factor, and then appending them to the list of HEK293 editing sites.

6.3 Per-site and per-transcript Endogenous RNA Age models (ERA)

6.3.1 Per-Site Age

Given a pre-determined editing rate, λ_i, and an observed fraction of edited reads (y_i), equation 6.2 was rearranged to estimate the mean age ,$\bar{\tau}_i$, of the population of transcripts containing a particular site i, giving

$$\bar{\tau}_i = -\frac{\ln(1 - y_i)}{\lambda_i}. \tag{6.4}$$

Here, equation 6.2 is rearranged rather than eq. 6.3 since the parameter, a_i, is only needed to account for the pre-existing level of editing at the start of the calibration experiment. Equation 6.4 calculates the mean time *since transcription* (i.e. 'age') not the time since ActD was added in the calibration experiment. To reflect this, $\bar{\tau}_i$ is used to represent 'age' to distinguish from experimental time variable t, and a_i is not needed.

In practice, some sites had no edited reads ($y_i = 0$) - in which case $\bar{\tau}_i = 0$ - whereas others were edited on all reads ($y_i = 1$) - in which case $\bar{\tau}_i = \infty$. To handle infinities, y_i was adjusted as

$$y_i' = \begin{cases} y_i - \dfrac{1}{R_i}, & y_i = 1, \\ \\ y_i, & \text{otherwise}, \end{cases} \tag{6.5}$$

where R_i is the number of reads mapping over site i.

The per-site age estimates $\bar{\tau}_i$ for the set of all sites annotated to a gene, $i \in I$ are assumed to be estimating the mean age of the same population of transcripts, and thus can be averaged to improve the precision of the estimate of the mean population age using the central limit theorem. Thus, the mean age of the population

of transcripts - or simply: the 'mean gene age' - is given by

$$\bar{\tau}_{\mathrm{g}} = \frac{1}{|I|} \sum_{i \in I} \bar{\tau}_i, \tag{6.6}$$

with standard deviation

$$\sigma_{\mathrm{g}} = \sqrt{\frac{1}{|I| - 1} \sum_{i \in I} \left(\bar{\tau}_i - \bar{\tau}_{\mathrm{g}}\right)^2}, \tag{6.7}$$

and a standard error of the mean

$$\mathrm{SEM}_{\mathrm{g}} = \frac{\sigma_{\mathrm{g}}}{\sqrt{|I|}}, \tag{6.8}$$

and - under the large-sample approximation - a 95% confidence interval (CI_{95}))

$$\mathrm{CI}_{95\%} = \bar{\tau}_{\mathrm{gene}} \pm 1.96 \times \mathrm{SEM}_{\mathrm{gene}}. \tag{6.9}$$

The performance of the per-site mean gene age estimates was evaluated by comparing the known duration of time between calibration time points with the change in the mean ages. For each gene in each of the calibration time points, the mean age of that gene in the 0h calibration time point was subtracted. A linear model was fit to these values using the `lm` function (as above) from which the gradient, $\hat{\beta}$ and the coefficient of correlation, R^2 was extracted.

6.3.2 Per-transcript Age

The age of a single transcript, τ_r, was modelled using a likelihood function for the age τ_r given the set of observed editing states at its sites, X:

$$L(\tau_r | X) = \prod_{i \in G} (1 - e^{-\lambda_i \tau_i}) \prod_{j \in A} e^{-\lambda_j \tau_j} \tag{6.10}$$

where $G = \{i : x_i = 1\}$ is the set of the edited sites, $A = \{j : x_j = 0\}$ is the set of the unedited sites and $X = G \cup A$.

The age τ_r is estimated by maximum likelihood estimation using the function the `maxLik` function (R package `maxLik` v1.5-2) on the log-likelihood function for each transcript, where the log-likelihood function is

$$\ell(\tau_r \mid X) = \sum_{i \in G} \ln\left(1 - e^{-\lambda_i \tau_r}\right) - \tau_r \sum_{j \in A} \lambda_j \tag{6.11}$$

and thus the MLE is given by

$$\hat{\tau}_r = \arg\max_{\tau} \ell(\tau_r \mid X). \tag{6.12}$$

The mean age of a population of transcripts, R, can be calculated from the MLEs as

$$\overline{\hat{\tau}}_{\mathrm{g}} = \frac{1}{|R|} \sum_{r \in R} \hat{\tau}_r. \tag{6.13}$$

For visualise purposes, this $\overline{\hat{\tau}}_{\mathrm{gene}}$ is typically denoted $\bar{\tau}_g$ and the individual MLEs are denoted as $\hat{\tau}_r$.

The distribution of the transcript age estimates for any given gene was represented as a histogram of the MLE values (typically between 0 and 20 hours with a bin width of 2 hours). For age analysis with no abundance data, the histograms were of frequencies (i.e. densities). To combine age and abundance into a single metric, the histograms were of counts and the the counts then normalised. Normalisation was done by calculating a scaling factor for each replicate or condition (depending on the setting) using the median of ratios method (Love et al., 2014) on the counts of reads per gene produced by IsoQuant (Prjibelski et al., 2022). Let $k_{g,s}$ be the raw count for gene g in sample s. The scaling factor was calculated as follows. The

geometric mean of counts for each gene g across all m samples is computed by

$$\mathrm{GM}_g = \left(\prod_{s=1}^{m} k_{g,s} \right)^{1/m}. \tag{6.14}$$

For each gene g and sample s, the ratio of its count to the gene's geometric mean is then calculated with

$$r_{g,s} = \frac{k_{g,s}}{\mathrm{GM}_g}. \tag{6.15}$$

The size factor for each sample s was defined as the median of these ratios over all genes:

$$s_s = \mathrm{median}_g\big(r_{g,s}\big) = \mathrm{median}_g\Big(\frac{k_{g,s}}{\mathrm{GM}_g}\Big). \tag{6.16}$$

These size factors are then used to normalise the histogram counts, to enable comparison of genes between samples.

In order to perform this analysis, the editing state of each site on each read must be extracted from the relevant .bam file containing the aligned reads. This is done using custom R code that implements the following algorithm which pulls out the base call and associated Q-score for every editing sites on every read mapped to a gene with fit sites:

Algorithm 2 RNA Editing Call Extraction Pipeline

Input: BAM file, IsoQuant read-to-gene mapping, editing rates
Output: The base calls at editing sites on a per-read basis
Define `chunk_size` as 10^6 reads
for each BAM file in `BAMS` **do**
    Create output directory if not exists
    Load IsoQuant read-to-gene mapping
    Filter reads to those mapping uniquely to genes containing fit sites
    Open BAM file with `chunk_size` yield
    Initialise `counter` to 1
    while BAM file has more reads to process **do**
        Read next chunk of BAM data
        Extract editing calls using `cigar_decoder_ranges_multi`:
            Parse CIGAR strings to determine read alignment structure
            Convert read alignments into genomic coordinate ranges
            Identify reads overlapping with specified editing sites
            Extract corresponding base calls from read sequences
            Extract corresponding quality-scores
            Store base calls and quality scores as strings in a data.frame object
        Save extracted calls to temporary file
        Increment `counter`
    end while
    Close BAM file
end for
Read all temporary files back into memory
Concatenate all extracted call data into a single dataset
if filtering on quality-score or editing rate is specified **then**
    Apply filtering criteria to concatenated dataset
end if
Save final processed dataset to output file

To avoid exhausting memory, the bam files were processed with Algorithm 2 in chunks of 5 million reads at a time. Likelihood functions are computed using function factories that are implemented in custom R code and the maximum likelihood estimates of each read that passes filtering determined as described above.

6.3.3 Differential Age

Differential age between two conditions was calculated by taking the $\log_2$ fold change of the mean age between conditions. Since the MLE values were not normally distributed, significance was determined using a non-exact Mann-Whitney U test with

tie-corrections on the MLE values, implemented with the `wilcox.test` function from the R package `stats`, v4.1.2. Replicates were pooled per-condition prior to testing. For each contrast (e.g. test condition vs control condition) tested, p-values were adjusted using the Benjamini-Hochberg method (implemented using the `p.adjust` function from the R `stats` package and results considered significant if p_{adjusted} was less than 0.05.

6.3.4 Calculation of half-lives

mRNA half-lives, $t_{1/2}$ were estimated in one of three ways.

1. From per-site mean ages:

$$t_{1/2} = \ln(2) \times \bar{\tau}_{gene} \tag{6.17}$$

2. From per-transcript mean age:

$$t_{1/2} = \ln(2) \times \overline{\hat{\tau}}_{gene} \tag{6.18}$$

3. By fitting

$$N(\hat{\tau}) = N_0 e^{-\delta\hat{\tau}} \tag{6.19}$$

to the distribution of MLEs using the non-linear least squares function `nls` from the R `stats` package. The half-life can then be found using the equation $t_{1/2} = \dfrac{\ln(2)}{\delta}$ with the value of δ obtained from the model fit.

The method for calculating half-lives described by Viegas et al. (2023) was implemented on here the NLambda-HEK calibration data as follows:

Transcript per-million (TPM) values were first averaged across replicates for each time point from the calibration experiment up to and including 12 hours post-ActD. Genes with fewer than six non-zero TPM measurements were excluded. A small

pseudocount $c = 10^{-4}$ was added to every TPM, and values were transformed to

$$\log_2 \mathrm{TPM}_g(t) = \log_2\big(\mathrm{TPM}_g(t) + c\big). \tag{6.20}$$

For each gene g, a linear regression was fit for the TPM values over the sequencing time points, t,

$$\log_2 \mathrm{TPM}_g(t) = \alpha_g + \beta_g\, t + \varepsilon, \tag{6.21}$$

and the coefficient of determination R^2, intercept α_g, and slope β_g extracted. A set of 100 'control' genes were selected by ranking those with

$$R^2 > 0.5, \quad \alpha_g > -3, \quad \beta_g > 0.02 \tag{6.22}$$

in descending order of β_g.

For each time point t, the mean TPM of the control genes

$$\mu_{\mathrm{ctrl}}(t) = \frac{1}{100} \sum_{g \in \mathrm{controls}} \log_2 \mathrm{TPM}_g(t) \tag{6.23}$$

was subtracted from the $\log_2$TPM values of all genes and a constant value of 2 added. This yield normalised values

$$\log_2 \mathrm{TPM}'_g(t) = \log_2 \mathrm{TPM}_g(t) - \mu_{\mathrm{ctrl}}(t) + 2. \tag{6.24}$$

A second linear fit

$$\log_2 \mathrm{TPM}'_g(t) = \alpha'_g + \beta'_g\, t + \varepsilon \tag{6.25}$$

provided the decay slope β'_g, of the abundances, from which the half-life was computed as

$$t_{1/2} = -\frac{1}{\beta'_g}. \tag{6.26}$$

Genes with $0 < t_{1/2} < 70$h were compared with half-life estimates from the ERA-derived methods detailed above.

6.4 Specific Analyses

6.4.1 Calibration

The calibration equation (Equation 6.3) was modified to account for the drop-off in editing observed in calibration time points far (greater than 12 hours) after ActD was added. The new equation,

$$y_i(t) = \beta_i(1 - a_i e^{-\lambda_i t}), \tag{6.27}$$

where the new site-specific parameter, β_i, was fit for each site i. The per-site mean age, $\bar{\tau}_i$ was thus given by

$$\bar{\tau}_i = -\frac{\ln(1 - \frac{y_i}{\beta_i})}{\lambda_i}. \tag{6.28}$$

6.4.2 Correlation Coefficients and visualisation

Pearson and Spearman correlation coefficients were calculated using the `cor.test` function from the R `stats` package. Plots of correlations (such as Figure 2.4D) were coloured by the two dimensional density of the data, which was calculated using the `kde2d` function from the R package `MASS` v7.3-58.3.

6.4.3 Q-score analysis

A Q-score (Phred Q-values) is calculated from an error probability, p_i as

$$Q_i = -10 \log_{10} p_i, \tag{6.29}$$

making Q-scores logarithmic. To calculate an accurate mean summary statistic, the Q-scores must first converted back to non-logarithmic values.

$$p_i = 10^{-Q_i/10}. \tag{6.30}$$

The mean error probability over a set of base calls, N, is then calculated as

$$\bar{p} = \frac{1}{N} \sum_{i=1}^{N} p_i, \tag{6.31}$$

and the corresponding mean Q-score score reported as

$$\bar{Q} = -10 \log_{10} \bar{p}. \tag{6.32}$$

Similarly, the median is calculated by converting to non-logarithmic values,

$$p_{(50)} = \text{median}\{\, p_i \,\}, \tag{6.33}$$

giving the median Q-score, $Q_{(50)}$ as

$$Q_{(50)} = -10 \log_{10} p_{(50)}. \tag{6.34}$$

6.4.4 Filtering transcript ages by number of states

Since calculating the mean age of a pool of transcripts, $\overline{\hat{\tau}}$ from a small number of MLE values, $\hat{\tau}$ is unreliable, the set of genes analysed was filtered based on the number of states observed. A 'state' is defined as a unique MLE value, and the default threshold used was to retain genes that 70 or more states across all of the samples in the experiment. An exception was made for the single cell analysis where a threshold of 10 or more states was used due to the low sequencing depth.

6.4.5 Differential Expression Analysis

Differentially expressed genes were determined using DESeq2 (Love et al., 2014). For short-read data, count matrices were obtained from Salmon v1.4.0 (Patro et al., 2017) whilst for long-read data they were obtained from IsoQuant v3.0.0 (Prjibelski et al., 2022). Standard arguments were used in both cases, p values were adjusted by the Benjmaini-Hochberg method with a significance level, α of 0.05 used. Shrunk

$\log_2$(fold change) values were used for the effect size and a threshold of $|\log_2(FC)| > 1$ used.

6.4.6 Gene Ontology analysis

Enrichment of Gene Ontology Biological Process (GO BP) terms was performed using the `enrichGO` function from the `clusterProfiler` R package with human annotation from `org.Hs.eg.db`. Analyses were carried out on Ensembl gene identifiers, specifying the ontology as 'BP' and using the Benjamini-Hochberg method for multiple-testing correction. For each contrast, the universe comprised all tested genes, and significantly differentially expressed genes were defined by an adjusted p-value below 0.05 and an absolute $\log_2$(Fold Change) of at least 1. For testing enrichment of differentially aged genes, the absolute $\log_2$(Fold Change) threshold was 0.5. GO BP terms with both adjusted $p < 0.05$ and $q < 0.05$ were considered enriched.

6.4.7 HEK293 Tet-on induction analysis

Details of the experimental procedure can be found in Appendix 6.4.9.

Due to the presence of short sequences mapping to the induced genes, sequencing reads were required to contain at least 10 fit sites, rather than 5. For each sample, transcript ages, $\hat{\tau}$ were estimated by MLE and mean gene ages, $\overline{\hat{\tau}}$ calculated as described above in Section 6.3. Transcript age distributions were calculated by creating frequency histograms for each sample individually. For visualisation purposes, each transcript age bin is represented as a mean across the replicates and with error bars denoting the standard error of the mean. The data are plotted in the middle of the bin on the x-axis.

Hierarchical clustering was performed using the `pheatmap` function from the R package `pheatmap` v1.0.12. Clustering was performed on the mean frequency values (as described in the paragraph above) using bins of transcript age with 2 hour widths,

between 0 hour and 20 hours.

Transcript abundance was quantified using IsoQuant v3.0.0, and for ease of visualization, transcript per thousand (TPK) values were calculated by taking the transcript per million (TPM) normalized output files and multiplying the values by 1,000.

6.4.8 Monocyte LPS Stimulation Analysis

Details of the experimental procedure can be found in Appendix 6.4.9.

Trajectory plots (for instance, Figure 4.6C) show the $\log_2$ fold change ($\log_2$FC) in both transcript age and counts for each replicate relative to the mean value of the 0 hour replicates and are plotted as points with a smoothed line (LOESS method), with the shaded area representing 95% confidence interval of the fit model.

The deviation of the 6 hour transcript age distributions from the expected distribution was calculated as follows: The distributions of the transcript ages in the 6 hour condition were represented as normalised count histograms (between 0 and 40 hours with a bin width of 2 hours) where the normalisation was done by the median of ratios method described in Methods Section 6.3.2. At this stage, genes for analysis were filtered to those that had a reasonable number of transcripts of old enough age to detect changes up to 6 hours in the past. A threshold of at 20 or more normalised counts in the 6-8 hour age bin was used.

Using the data from the 0 hour time point (i.e. prior to LPS stimulation) the half-life model of equation 6.19 was then fit and the decay rate, δ, stored.

Then to generate the expected distribution of transcript ages for the 6 hour condition, the equation

$$N(\hat{\tau}) = N_0 e^{-\delta\hat{\tau}} \tag{6.35}$$

was used where N_0 was the normalised counts of transcripts in the first age bin (0-2 hours) from the 6 hour condition and δ was the decay rate from the 0 hour condition. To calculate the deviation of the transcript ages from the expected distribution, the

expected values were subtracted from the normalised counts for each histogram bin.

The resulting values for each gene were Z-score scaled (the mean subtracted and then divided by the standard deviation) and clustered by k-means clustering using $k = 10$ and the `kmeans` function from the R `stats` package.

6.4.9 Single-cell cell cycle analysis

Transcript ages in single cells were estimated as above. The labelling of cells as G1/S or G2/M was done by my collaborator Aaron, who ranked the cells based on the ratio of the TPM values for the G1/S and G2/M genes identified by Tirosh et al. (2016). The top 15 cells of both categories were selected, with the rest of the cells being labelled as 'neither' G1/S nor G2/M.

Transcript age distributions were calculated for the only two cell-cycle associated genes that were present in the data with sufficient sequencing depth: HELLS and TMPO. Transcript age distributions were created for each of the three cell labels (G1/S, G2/M, Neither) by pooling the individual transcript MLEs, $\hat{\tau}$ for the cells in each category. Frequency histograms were generated over the age range 0 to 20 hours with a bin width of 4 hours.

Appendix 1: Wet lab Methods

6.5 Cell Culture

6.5.1 HEK293

HEK293-FT cells were cultured in DMEM (Thermo Fisher, 11965092) supplemented with 10% FBS (Gibco) and 1% Penicillin-streptomycin (Sigma). HEK293T cells were seeded on tissue culture plastic 24-well plates in triplicate 48 hours before the experiment.

6.5.2 Human hiPSC Derived Neurons

Human induced pluripotent stem cells (hiPSCs) were generated from reprogrammed fibroblasts from healthy donors, with approval from the London-Hamstead Research Ethics Committee and the University College London, Great Ormond Street Institute of Child Health and Great Ormond Street Hospital Joint Research Office. The hiPSCs were cultured on Geltrex (Thermo Fisher) in E8 media (Thermo Fisher) or mTeSR (Stem Cell Technologies) and passaged using 0.5 mM ethylenediaminetetraacetic acid (Thermo Fisher). Neurons were generated using an established protocol Shi et al. (2012). Briefly, neocortical stem cell differentiation was achieved with dual SMAD inhibition using SB431542 (10 μM, Tocris) and dorsomorphin dihydrochloride (1 μM, Tocris) for 12 days, followed by ongoing culture for an additional 90 days to differentiate into neurons.

6.5.3 Human Organotypic Brain Slices

Organotypic brain slice cultures were prepared using the interface method adapted from Ravi et al. (2019) and De Simoni and Yu (2006). Fresh healthy cortical human brain tissue was removed during neurosurgery and used for research under appro-

priate National Research Ethics approval (Reference: 21/SC/0111). The tissue was resected and placed in ice-cold dissection media in the operating theatre and transferred for immediate sectioning at 300 μM thickness using a Leica VT1200s vibratome. Each 1 cm^2 brain section was plated on a 30 mm culture plate insert (Millipore, PICM03050) in a 35 mm Nunc 6-well plate flooded with 1 ml of culture media per well and transferred to an incubator at 37° C and 5% CO_2. Media was changed every 24 hours.

Dissection media was composed of Hank's balanced salt solution (HBSS) supplemented with HEPES (pH 7.4, 2.5 mM), D-glucose (30 mM), $CaCl_2$ (1 mM), $MgCl_2$ (1 mM), and $NaHCO_3$ (4 mM). Culture media was composed of Neurobasal L-Glutamine supplemented with 2% serum-free B-27, 2% Anti-Anti, 13 mM D-glucose, 1 mM $MgSO_4$,15 mM HEPES, and 2 mM GlutaMAX. RNA extractions were performed in triplicate using the Qiagen RNeasy Plus Universal Mini Kit, following manufacturer protocols.

6.6 Organoid Culture

Cultures of human intestinal organoids were established as previously described in Sato et al. (2011). Briefly, organoids were re-established in culture and maintained as three-dimensional spheroids in Cultrex^{TM} reduced growth factor basement membrane extract, type 2 (R&D Systems) from previously cryopreserved organoids originally derived from human intestinal biopsies Meran et al. (2020) (Research Ethics Committee references 04-Q0508-79 and 18/EE/0150). Organoids were cultured in 24-well plates using human IntestiCult Organoid Growth Medium (STEMCELL Technologies) supplemented with 3 μM CHIR99021 during the expansion phase, with a passaging ratio of 1:5 and media changes every 2 days.

For hypoxia experiments, organoids were treated with 1 mM Dimethyloxallyl Glycine (DMOG) for 6 hours. For necroptosis experiments, organoids were stimulated with 30 ng/ml of recombinant TNF-alpha (R&D Systems) for 6 hours. RNA extractions

were performed in triplicate using the Qiagen RNeasy Plus Mini Kit, following manufacturer protocols.

6.7 Calibration Protocol

Cells (either HEK293s or hiPSCs) were cultured as previously detailed. Actinomycin D was added to the cells to a final concentration of 1 μg/ml in complete media at time $t = 0$ hours. Wells were lysed in triplicate at timed intervals. For example, the HEK293 cells were lysed at the following timepoints: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 16, 24, 38 hours.

RNA was extracted using the RNeasy Plus Mini Kit (Qiagen 74136) as per the manufacturer's instructions. Cells were lysed in the wells using RLT Plus from the Qiagen kit and vortexed to homogenise. Purified total RNA was then quantified using the Qubit fluorometer (Life Technologies Q32855), and quality was measured on an Agilent Tapestation using RNA Screentape (Agilent 5067-5576).

Full-length, short-read sequencing libraries for Illumina were then prepared by poly-A selection using oligo-dT magnetic beads (NEB E7490) followed by the Ultra II directional library prep kit for Illumina (NEB E7760) according to the manufacturer's instructions. In brief, RNA was fragmented, a library was prepared by reverse transcription with random primers, and adapters were ligated to both ends of the cDNA. Uracil was incorporated during second-strand synthesis, and strand specificity was achieved by digestion of the second strand by USER. Sequencing adapters and sample barcodes were added by PCR before purification, QC, and pooling. Short-read sequencing was performed on a NovaSeq (Illumina) using 150 cycle kits with 76 bps for read 1 and 2. The samples were sequenced at a depth of approximately 225 million reads per timepoint (HEK293) and 500 million reads per timepoint (Cortex).

Long-read sequencing of the HEK293 and Cortex samples was performed using PacBio Iso-Seq on a Sequel IIe. Library prep used the Iso-Seq Express Template

Preparation kit (PacBio) according to the manufacturer's instructions. Both time-points in the Cortex calibration were sequenced on one Sequel IIe SMRT Cell 8M flow cell in CCS mode at a mean depth of approximately 2.2M reads. In subsequent long-read sequencing experiments, the Oxford Nanopore Technologies platform was utilised given its higher throughput, with similar performance in determining single-molecule ages.

6.8 Human Gene Induction Experiment

6.8.1 Cloning

Total RNA was extracted from HEK293 cells using the RNeasy Mini Kit, and a cDNA library was generated. *ATG14*, *BVES*, and *ACBD7* were then amplified from this cDNA library using transcript-specific primers. Bands were then gel-purified and assembled into a plasmid vector using NEBuilder Hi-Fi DNA Assembly (E2621). *E. cloni* 10G Chemically Competent Cells (Lucigen) were transformed and grown overnight.

The sequencing data allowed us to identify pre-edited ADAR editing sites. PCR was used to restore these editing sites to their unedited form. The PCR products from these reactions were then reassembled using NEBuilder Hi-Fi assembly and validated using whole-plasmid sequencing. The final products were plasmids, each containing the full cDNAs of the genes *ATG14*, *BVES*, and *ACBD7*, respectively. These plasmids are known henceforth as pAG051, pAG053, and pAG054, respectively.

6.8.2 Cell Transfection

HEK293 cells were transfected using the Transit-X2 delivery system (Mirus Bio) when the cells reached 70% confluency. Each well of a 24-well plate received 250 ng of tTA3 transactivator plasmid and 250 ng of an equimolar mixture of the three mammalian cDNA plasmids (pAG051, pAG053, and pAG054). Transfection was

performed 48 hours before the $t = 0$ timepoint.

The ON, OFF, and PULSE conditions were performed as follows:

- **ON:** At $t = 0$ hours, doxycycline was added. Wells were lysed at 0, 0.5, 1, 4, and 8 hours. A mock induction was lysed at 8 hours.
- **OFF:** Doxycycline was added 24 hours before the $t = 0$ timepoint. Cells were changed to doxycycline-free media at $t = 0$ hours and then lysed at $t = 16$ hours.
- **PULSE:** At $t = 0$ hours, doxycycline was added. At $t = 8$ and $t = 16$ hours, the cells were lysed.

For each of these conditions, doxycycline was added to a concentration of 1 μg/ml. Wells were lysed in triplicate at each timepoint.

6.9 Library Preparation and Sequencing

Library preparation was undertaken using a bespoke SMARTseq protocol designed to minimise the number of PCR cycles to reduce PCR errors (courtesy of Adam Cribbs and Danson Loi). In brief, at least 100 ng of total RNA was denatured at 65° C in the presence of 100 μM poly-T reverse primer and 1 mM dNTPs. Reverse transcription was performed by adding RT buffer (Thermo Fisher), Maxima H-reverse transcriptase (Thermo Fisher), and a template-switching oligonucleotide. A bespoke temperature ramp-up and cycle programme was used.

Excess primers were digested by incubation with heat-labile Exonuclease I (Thermo Fisher). Sample clean-up was performed with SPRI-select beads (0.6X by volume). Initial PCR was performed with 2X KAPA ReadyMix (Roche) and ONT PCR handle primer. This was followed by a second PCR, with the reaction split into four.

Samples were prepared for sequencing using the ligation sequencing kit v.14 (ONT, SQK-LSK-114) as per the manufacturer's instructions. Samples were sequenced on

a PromethION24 sequencer (ONT) over five R10.4.1 flow cells, yielding an average of approximately 70 Gb per flow cell.

6.10 Monocyte LPS Stimulation Experiment

6.10.1 Peripheral Blood Mononuclear Cell (PBMC) Isolation

Whole blood was sourced from a healthy male via the UK National Health Service blood bank. Blood volume was standardized by supplementing with Hank's Balanced Salt Solution (HBSS) containing 0.3 mM EDTA to a final volume of 25 ml.

Next, 20 ml of Ficoll-Paque was dispensed into separate 50 ml tubes. The blood was delicately layered atop the Ficoll using a 5 ml pipette, ensuring the preservation of the discrete interface. The sample was centrifuged at 700 g, without brake, for 25 minutes at room temperature.

Post-centrifugation, the PBMCs, identifiable as a white interphase layer, were harvested using a Pasteur pipette and transferred into a clean 50 ml tube. The cell suspension was diluted to 50 ml with HBSS 0.3 mM EDTA, followed by centrifugation at 500 g for 10 minutes at room temperature. The supernatant was decanted, and the cell pellet was subsequently resuspended in 40 ml HBSS 0.3 mM EDTA. This wash step was repeated once. After the final centrifugation, the PBMCs were resuspended in 1 ml of MACS buffer in preparation for CD14+ isolation.

6.10.2 CD14+ Cell Isolation

Cells were processed using the MagniSort system (Invitrogen). Initially, 200 μl of MagniSort Enrichment Antibody was added and thoroughly vortexed. The mixture was incubated at room temperature (RT) for 10 minutes.

Subsequently, 3 ml of MACS buffer was introduced, followed by centrifugation at 300 g for 3 minutes at RT. The pellet was resuspended in 1 ml MACS buffer. Next, 300

μl of MagniSort Positive Selection Beads was added, and the solution was vortexed. Following another 10-minute incubation at RT, the volume was adjusted to 2.5 ml with MACS buffer and mixed by pipetting.

The tube was placed in a magnet and incubated at RT for 5 minutes. The supernatant, containing CD14- cells, was discarded, and the magnetic separation was repeated twice more after resuspending the beads in MACS buffer. The retained CD14+ cells were supplemented with 30 ml Iscove's Modified Dulbecco's Medium containing 10% Fetal Bovine Serum, 1% Penicillin-Streptomycin-Glutamine, and x μM GM-CSF.

6.10.3 LPS Induction

Human CD14+ monocytes were plated in triplicate in 6-well plates at a density of 2 million cells per well immediately after magnetic isolation and 24 hours before LPS stimulation. At $t = 0$ hours, *E. coli* lipopolysaccharide (LPS), purified by ion-exchange chromatography (Sigma L3024), was added to each experimental well to a final concentration of 100 ng/ml.

6.10.4 Library Preparation and Sequencing

Samples were reverse transcribed and amplified using the same protocol as with the HEK293 induction experiment (see above). They were sequenced on a PromethION24 machine (ONT) over eighteen R10.4.1 flow cells (one per sample) at an average depth of approximately 70M reads each, yielding over 1.2 billion long reads.

6.11 Single-Cell RNA Sequencing in Plates

HEK293FT cells (Thermo Fisher) were grown to 70% confluency, trypsinised, counted, and resuspended in FACS buffer (PBS, 2% molecular grade BSA, 1 μg/ml actinomycin D) to a final concentration of 5×10^6 cells per ml. Shortly before cell sorting, DAPI was added as a viability dye.

Before the first sort, the FACS machine was calibrated using a horseradish peroxidase assay to ensure that cells were deposited into each well. Single cells were sorted into single wells of a low-bind 96-well plate (Eppendorf) using the MoFlo XDP cell sorter (Beckman Coulter) with a 100 μm nozzle and the following lasers and filters: 405 nm, 447/60. After gating on FSC/SSC and a doublet gate, DAPI-negative cells were selected for sorting.

The NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB E6421L) was used to prepare SMARTseq libraries. Cells were sorted into wells containing 5 μl of fresh lysis buffer. Immediately after sorting, samples were prepared for reverse transcription and library preparation.

After reverse transcription and 22 cycles of PCR amplification as per the manufacturer's instructions, an additional 5 cycles of PCR amplification were undertaken to add nanopore barcodes and adaptors. Samples were then pooled and prepared for sequencing using the ligation sequencing kit v.14 (ONT, SQK-LSK-114) as per the manufacturer's instructions.

The library was sequenced on a PromethION24 machine (ONT) over six R10.4.1 flow cells with a mean output of 44M reads per flow cell.

6.12 Hyperactive Editor Experiments

HEK293 cells were transfected with the editors (and corresponding guide RNAs if required). Twenty-four hours later, actinomycin-D was added to stop transcription, and cells were lysed at 0-hour and 8-hour timepoints in triplicate.

In the first generation of editor selection, we studied the full-length hyperactive ADAR2 mutant ADAR2(E488Q). Additionally, several fusion proteins of the ADAR2 (E488Q) catalytic domain were tested: ADAR2(E488Q)-Nlambda (Montiel-Gonźalez et al., 2016), SNAP-ADAR2(E488Q) (Vogel et al., 2018), Cas13-ADAR2(E488Q) (Cox et al., 2017), and PABP-ADAR2(E488Q) (courtesy of Rory Maizels). We also

tested ABEmax (Matsoukas, 2018) and the cytosine base editor BE3 (Grünewald et al., 2019). Although SNAP-ADAR2(E488Q) showed slightly higher levels of editing than ADAR2(E488Q)-Nlambda, we decided not to proceed with it because its guide RNA cannot be genetically encoded due to the need for an O6-benzylguanine base.

For the second generation of editors, we used a chassis based on the *E. coli* tRNA-specific adenosine deaminase (TadA). Plasmids for ADAR2(E488Q)-Nlambda (the best construct from generation 1), TadA7.10 (Gaudelli et al., 2020), TadA8.20 (Xiao et al., 2023), TadA8.20-Nlambda, and TadA8.20-Dps-N (Kwon and Giessen, 2022; Park et al., 2020) (Dps-N construct courtesy of Karl Brune) were transfected into HEK293 cells and the experiment performed as for the first generation editors.

Bibliography

Agarwal, V. and Kelley, D. R. (2022), 'The genetic and biochemical determinants of mRNA degradation rates in mammals', *Genome Biology* **23**(1), 1–28.

Al Khafaji, A. M., Smith, J. T., Garimella, K. V., Babadi, M., Popic, V., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M. A., Blaum, E. M., Day, A., Costello, M., Bowers, T., Gabriel, S., Banks, E., Philippakis, A. A., Boland, G. M., Blainey, P. C. and Hacohen, N. (2023), 'High-throughput RNA isoform sequencing using programmed cDNA concatenation', *Nature Biotechnology 2023 42:4* **42**(4), 582–586.

Alpert, T., Herzel, L. and Neugebauer, K. M. (2016), 'Perfect timing: splicing and transcription rates in living cells', *Wiley interdisciplinary reviews. RNA* **8**(2), 10.1002/wrna.1401.

Arakawa, M., Uriu, K., Saito, K., Hirose, M., Katoh, K., Asano, K., Nakane, A., Saitoh, T., Yoshimori, T., Morita, E. and Hurley, J. H. (2025), 'HEATR3 recognizes membrane rupture and facilitates xenophagy in response to Salmonella invasion', *Proceedings of the National Academy of Sciences* **122**(14), e2420544122.

Arribere, J. A. and Fire, A. Z. (2018), 'Nonsense mRNA suppression via nonstop decay', *eLife* **7**, e33292.

Barba, M., Czosnek, H. and Hadidi, A. (2014), 'Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology', *Viruses* **6**(1), 106.

Barenco, M., Brewer, D., Papouli, E., Tomescu, D., Callard, R., Stark, J. and Hubank, M. (2009), 'Dissection of a complex transcriptional response using genome-wide transcriptional modelling', *Molecular Systems Biology* **5**, 327.

Baron-Benhamou, J., Gehring, N. H., Kulozik, A. E. and Hentze, M. W. (2004), 'Using the λN Peptide to Tether Proteins to RNAs', *mRNA Processing and Metabolism* pp. 135–153.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F. J., Rechavi, G., Li, J. B., Eisenberg, E. and Levanon, E. Y. (2014), 'A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes', *Genome Research* **24**(3), 365.

Ben-Ari, Y., Brody, Y., Kinor, N., Mor, A., Tsukamoto, T., Spector, D. L., Singer, R. H. and Shav-Tal, Y. (2010), 'The life of an mRNA in space and time', *Journal of Cell Science* **123**(10), 1761.

Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)* **57**(1), 289–300.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell,

J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Catenazzi, M. C. E., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008), 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature 2008 456:7218* **456**(7218), 53–59.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A. and Theis, F. J. (2020), 'Generalizing RNA velocity to transient cell states through dynamical modeling', *Nature Biotechnology 2020 38:12* **38**(12), 1408–1414.

Bhattarai-Kline, S., Lear, S. K., Fishman, C. B., Lopez, S. C., Lockshin, E. R., Schubert, M. G., Nivala, J., Church, G. M. and Shipman, S. L. (2022), 'Recording gene expression order in DNA by CRISPR addition of retron barcodes', *Nature 2022 608:7921* **608**(7921), 217–225.

Boo, S. H. and Kim, Y. K. (2020), 'The emerging role of RNA modifications in the regulation of mRNA stability', *Experimental & Molecular Medicine 2020 52:3* **52**(3), 400–408.

Braun, K. A. and Young, E. T. (2014), 'Coupling mRNA Synthesis and Decay', *Molecular and Cellular Biology* **34**(22), 4078–4087.

Brennan, C. M. and Steitz, J. A. (2001), 'HuR and mRNA stability', *Cellular and Molecular Life Sciences: CMLS* **58**(2), 266.

Brun, N. R., Salanga, M. C., Mora-Zamorano, F. X., Lamb, D. C., Goldstone, J. V. and Stegeman, J. J. (2021), 'Orphan cytochrome P450 20a1 CRISPR/Cas9 mutants and neurobehavioral phenotypes in zebrafish', *Scientific Reports 2021 11:1* **11**(1), 1–11.

Bushnell, D. A., Cramer, P. and Kornberg, R. D. (2002), 'Structural basis of transcription: α-AmanitinâRNA polymerase II cocrystal at 2.8 â« resolution', *Proceedings of the National Academy of Sciences of the United States of America* **99**(3), 1218.

Chan, T. W., Fu, T., Bahn, J. H., Jun, H. I., Lee, J. H., Quinones-Valdez, G., Cheng, C. and Xiao, X. (2020), 'RNA editing in cancer impacts mRNA abundance in immune response pathways', *Genome Biology* **21**(1), 1–25.

Chang, L., Zhang, Z., Yang, J., McLaughlin, S. H. and Barford, D. (2014), 'Molecular architecture and mechanism of the anaphase-promoting complex', *Nature 2014 513:7518* **513**(7518), 388–393.

Chen, C. Y. A. and Shyu, A. B. (1995), 'AU-rich elements: characterization and importance in mRNA degradation', *Trends in biochemical sciences* **20**(11), 465–470.

Chen, H., Liu, X. and Patel, D. J. (1996), 'DNA bending and unwinding associated with actinomycin D antibiotics bound to partially overlapping sites on DNA', *Journal of molecular biology* **258**(3), 457–479.

Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A. and Deplancke, B. (2022), 'Live-seq enables temporal transcriptomic recording of single cells', *Nature 2022 608:7924* **608**(7924), 733–740.

Chung, H., Calis, J. J., Wu, X., Sun, T., Yu, Y., Sarbanes, S. L., Dao Thi, V. L., Shilvock, A. R., Hoffmann, H. H., Rosenberg, B. R. and Rice, C. M. (2018), 'Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown', *Cell* **172**(4), 811–824.

Clark, T. A., Schweitzer, A. C., Chen, T. X., Staples, M. K., Lu, G., Wang, H., Williams, A. and Blume, J. E. (2007), 'Discovery of tissue-specific exons using comprehensive human exon microarrays', *Genome Biology* **8**(4), 1–16.

Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R. and Boothroyd, J. C. (2005), 'Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay', *Nature Biotechnology 2005 23:2* **23**(2), 232–237.

Cox, D. B., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J., Joung, J. and Zhang, F. (2017), 'RNA editing with CRISPR-Cas13', *Science* **358**(6366), 1019–1027.

Crick, F. (1970), 'Central Dogma of Molecular Biology', *Nature 1970 227:5258* **227**(5258), 561–563.

De Coster, W. and Rademakers, R. (2023), 'NanoPack2: population-scale evaluation of long-read sequencing data', *Bioinformatics* **39**(5).

De Jonghe, J., Opzoomer, J. W., Vilas-Zornoza, A., Nilges, B. S., Crane, P., Vicari, M., Lee, H., Lara-Astiaso, D., Gross, T., Morf, J., Schneider, K., Cudini, J., Ramos-Mucci, L., Mooijman, D., Tiklová, K., Salas, S. M., Langseth, C. M., Kashikar, N. D., Carrami, E. M., McIntyre, R., Swerner, C. B., Hessel, E. M., Kapourani, C. I. A., Regep, C., Roberts, C. E., Schapiro, D., Lundeberg, J., Nilsson, M., Shalek, A. K., Cribbs, A. P. and Taylor-King, J. P. (2024), 'scTrends: A living review of commercial single-cell and spatial 'omic technologies', *Cell Genomics* **4**(12), 100723.

de Reuver, R., Verdonck, S., Dierick, E., Nemegeer, J., Hessmann, E., Ahmad, S., Jans, M., Blancke, G., Van Nieuwerburgh, F., Botzki, A., Vereecke, L., van Loo, G., Declercq, W., Hur, S., Vandenabeele, P. and Maelfait, J. (2022), 'ADAR1 prevents autoinflammation by suppressing spontaneous ZBP1 activation', *Nature 2022 607:7920* **607**(7920), 784–789.

De Simoni, A. and Yu, L. M. (2006), 'Preparation of organotypic hippocampal slice cultures: interface method', *Nature Protocols 2006 1:3* **1**(3), 1439–1445.

Delaunay, S., Helm, M. and Frye, M. (2023), 'RNA modifications in physiology and disease: towards clinical applications', *Nature Reviews Genetics 2023 25:2* **25**(2), 104–122.

Deng, X., Sun, L., Zhang, M., Basavaraj, R., Wang, J., Weng, Y. L. and Gao, Y. (2025), 'Biochemical profiling and structural basis of ADAR1-mediated RNA editing', *Molecular Cell* **85**(7), 1381–1394.

Desterro, J. M., Keegan, L. P., Jaffray, E., Hay, R. T., O'Connell, M. A. and Carmo-Fonseca, M. (2005), 'SUMO-1 modification alters ADAR1 editing activity', *Molecular biology of the cell* **16**(11), 5115–5126.

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P. and Koszinowski, U. H. (2008), 'High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay', *RNA* **14**(9), 1959–1972.

D'Orazio, K. N. and Green, R. (2021), 'Ribosome states signal RNA quality control', *Molecular Cell* **81**(7), 1372–1383.

D'Orazio, K. N., Wu, C. C.-C., Sinha, N., Loll-Krippleber, R., Brown, G. W. and Green, R. (2019), 'The endonuclease Cue2 cleaves mRNAs at stalled ribosomes during No Go Decay', *eLife* **8**.

Dowdle, M. E. and Lykke-Andersen, J. (2025), 'Cytoplasmic mRNA decay and quality control machineries in eukaryotes', *Nature Reviews Genetics 2025* pp. 1–16.

Finkel, Y., Gluck, A., Nachshon, A., Winkler, R., Fisher, T., Rozman, B., Mizrahi, O., Lubelsky, Y., Zuckerman, B., Slobodin, B., Yahalom-Ronen, Y., Tamir, H.,

Ulitsky, I., Israely, T., Paran, N., Schwartz, M. and Stern-Ginossar, N. (2021), 'SARS-CoV-2 uses a multipronged strategy to impede host protein synthesis', *Nature 2021 594:7862* **594**(7862), 240–245.

Gal-Ben-Ari, S., Barrera, I., Ehrlich, M. and Rosenblum, K. (2019), 'PKR: A kinase to remember', *Frontiers in Molecular Neuroscience* **11**, 425641.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I. and Liu, D. R. (2017), 'Programmable base editing of Aâ¢T to Gâ¢C in genomic DNA without DNA cleavage', *Nature 2017 551:7681* **551**(7681), 464–471.

Gaudelli, N. M., Lam, D. K., Rees, H. A., Solá-Esteves, N. M., Barrera, L. A., Born, D. A., Edwards, A., Gehrke, J. M., Lee, S. J., Liquori, A. J., Murray, R., Packer, M. S., Rinaldi, C., Slaymaker, I. M., Yen, J., Young, L. E. and Ciaramella, G. (2020), 'Directed evolution of adenine base editors with increased activity and therapeutic application', *Nature biotechnology* **38**(7), 892–900.

George, C. X. and Samuel, C. E. (1999), 'Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible', *Proceedings of the National Academy of Sciences of the United States of America* **96**(8), 4621–4626.

Gładysz, M., Andrałojć, W., Czapik, T., Gdaniec, Z. and Kierzek, R. (2019), 'Thermodynamic and structural contributions of the 6-thioguanosine residue to helical properties of RNA', *Scientific Reports 2019 9:1* **9**(1), 1–8.

Grünewald, J., Zhou, R., Iyer, S., Lareau, C. A., Garcia, S. P., Aryee, M. J. and Joung, J. K. (2019), 'CRISPR DNA base editors with reduced RNA off-target and self-editing activities', *Nature biotechnology* **37**(9), 1041–1048.

Guha, M. and Mackman, N. (2001), 'LPS induction of gene expression in human monocytes', *Cellular Signalling* **13**(2), 85–94.

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G. J., Larsson, A. J., Faridani, O. R. and Sandberg, R. (2020), 'Single-cell RNA counting at allele and isoform resolution using Smart-seq3', *Nature Biotechnology 2020 38:6* **38**(6), 708–714.

Hartenian, E. and Glaunsinger, B. A. (2019), 'Feedback to the central dogma: cytoplasmic mRNA decay and transcription are interdependent processes', *Critical Reviews in Biochemistry and Molecular Biology* **54**(4), 385–398.

Herbert, A., Alfken, J., Kim, Y. G., Mian, I. S., Nishikura, K. and Rich, A. (1997), 'A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase', *Proceedings of the National Academy of Sciences of the United States of America* **94**(16), 8421–8426.

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., Von Haeseler, A., Zuber, J. and Ameres, S. L. (2017), 'Thiol-linked alkylation of RNA to assess expression dynamics', *Nature Methods 2017 14:12* **14**(12), 1198–1204.

Higuchi, M., Single, F. N., Köhler, M., Sommer, B., Sprengel, R. and Seeburg, P. H. (1993), 'RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency', *Cell* **75**(7), 1361–1370.

Hir, H. L., Saulière, J. and Wang, Z. (2015), 'The exon junction complex as a node of post-transcriptional networks', *Nature Reviews Molecular Cell Biology 2015 17:1* **17**(1), 41–54.

Hu, S. B., Heraud-Farlow, J., Sun, T., Liang, Z., Goradia, A., Taylor, S., Walkley, C. R. and Li, J. B. (2023), 'ADAR1p150 prevents MDA5 and PKR activation via distinct mechanisms to avert fatal autoinflammation', *Molecular Cell* **83**(21), 3869–3884.

Imamachi, N., Tani, H., Mizutani, R., Imamura, K., Irie, T., Suzuki, Y. and Akimitsu, N. (2014), 'BRIC-seq: A genome-wide approach for determining RNA stability in mammalian cells', *Methods* **67**(1), 55–63.

Jonkers, I. and Lis, J. T. (2015), 'Getting up to speed with transcription elongation by RNA polymerase II', *Nature Reviews Molecular Cell Biology 2015 16:3* **16**(3), 167–177.

Jürges, C., Dölken, L. and Erhard, F. (2018), 'Dissecting newly transcribed and old RNA using GRAND-SLAM', *Bioinformatics* **34**(13), i218–i226.

Kakihana, A., Oto, Y., Saito, Y. and Nakayama, Y. (2019), 'Heat shock-induced mitotic arrest requires heat shock protein 105 for the activation of spindle assembly checkpoint', *The FASEB Journal* **33**(3), 3936–3953.

Karki, R., Sundaram, B., Sharma, B. R., Lee, S. J., Malireddi, R. K., Nguyen, L. N., Christgen, S., Zheng, M., Wang, Y., Samir, P., Neale, G., Vogel, P. and Kanneganti, T. D. (2021), 'ADAR1 restricts ZBP1-mediated immune response and PANoptosis to promote tumorigenesis', *Cell Reports* **37**(3).

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W. (2015), 'Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells', *Cell* **161**(5), 1187–1201.

Kögel, A., Keidel, A., Loukeri, M. J., Kuhn, C. C., Langer, L. M., Schäfer, I. B. and Conti, E. (2024), 'Structural basis of mRNA decay by the human exosomeâribosome supercomplex', *Nature 2024 635:8037* **635**(8037), 237–242.

Kratochvill, F., Gratz, N., Qualls, J. E., Van De Velde, L. A., Chi, H., Kovarik, P. and Murray, P. J. (2015), 'Tristetraprolin limits inflammatory cytokine production in tumor-associated macrophages in an mRNA decay-independent manner', *Cancer Research* **75**(15), 3054–3064.

Kumar, R., Khan, M., Francis, V., Aguila, A., Kulasekaran, G., Banks, E. and McPherson, P. S. (2024), 'DENND6A links Arl8b to a Rab34/RILP/dynein complex, regulating lysosomal positioning and autophagy', *Nature Communications 2024 15:1* **15**(1), 1–18.

Kwon, S. and Giessen, T. W. (2022), 'Engineered Protein Nanocages for Concurrent RNA and Protein Packaging In Vivo', *ACS synthetic biology* **11**(10), 3504.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S. and Kharchenko, P. V. (2018), 'RNA velocity of single cells', *Nature 2018 560:7719* **560**(7719), 494–498.

Lebrigand, K., Magnone, V., Barbry, P. and Waldmann, R. (2020), 'High throughput error corrected Nanopore single cell transcriptome sequencing', *Nature Communications 2020 11:1* **11**(1), 1–8.

Lee, K., Ku, J., Ku, D. and Kim, Y. (2024), 'Inverted Alu repeats: friends or foes in the human transcriptome', *Experimental & Molecular Medicine 2024 56:6* **56**(6), 1250–1262.

Li, Y., Junod, S. L., Ruba, A., Kelich, J. M. and Yang, W. (2018), 'Nuclear export of mRNA molecules studied by SPEED microscopy', *Methods (San Diego, Calif.)* **153**, 46.

Liddicoat, B. J., Piskol, R., Chalk, A. M., Ramaswami, G., Higuchi, M., Hartner, J. C., Li, J. B., Seeburg, P. H. and Walkley, C. R. (2015), 'RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself', *Science* **349**(6252), 1115–1120.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008), 'Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis', *Cell* **133**(3), 523–536.

Love, M. I., Huber, W. and Anders, S. (2014), 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology* **15**(12), 1–21.

Lugowski, A., Nicholson, B. and Rissland, O. S. (2018), 'Determining mRNA half-lives on a transcriptome-wide scale', *Methods* **137**, 90–98.

Lykke-Andersen, S., Piñol-Roma, S. and Kjems, J. (2007), 'Alternative splicing of the ADAR1 transcript in a region that functions either as a 5â^2 – $UTRoranORF', RNA$**13**(10), 1732.

MacKenzie, M. and Argyropoulos, C. (2023), 'An Introduction to Nanopore Sequencing: Past, Present, and Future Considerations', *Micromachines 2023, Vol. 14, Page 459* **14**(2), 459.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. and McCarroll, S. A. (2015), 'Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets', *Cell* **161**(5), 1202–1214.

Matsoukas, I. G. (2018), 'Commentary: Programmable base editing of A·T to G·C in genomic DNA without DNA cleavage', *Frontiers in Genetics* **9**(FEB), 21.

Mboukou, A., Rajendra, V., Messmer, S., Mandl, T. C., Catala, M., Tisné, C., Jantsch, M. F. and Barraud, P. (2024), 'Dimerization of ADAR1 modulates site-specificity of RNA editing', *Nature Communications 2024 15:1* **15**(1), 1–14.

Melé, M., Mattioli, K., Mallard, W., Shechner, D. M., Gerhardinger, C. and Rinn, J. L. (2017), 'Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs', *Genome research* **27**(1), 27–37.

Meran, L., Massie, I., Campinoti, S., Weston, A. E., Gaifulina, R., Tullie, L., Faull, P., Orford, M., Kucharska, A., Baulies, A., Novellasdemunt, L., Angelis, N., Hirst, E., König, J., Tedeschi, A. M., Pellegata, A. F., Eli, S., Snijders, A. P., Collinson, L., Thapar, N., Thomas, G. M., Eaton, S., Bonfanti, P., De Coppi, P. and Li, V. S. (2020), 'Engineering transplantable jejunal mucosal grafts using patient-derived organoids from children with intestinal failure', *Nature Medicine* **26**(10), 1593–1601.

Mladenova, D., Barry, G., Konen, L. M., Pineda, S. S., Guennewig, B., Avesson, L., Zinn, R., Schonrock, N., Bitar, M., Jonkhout, N., Crumlish, L., Kaczorowski, D. C., Gong, A., Pinese, M., Franco, G. R., Walkley, C. R., Vissel, B. and Mattick, J. S. (2018), 'Adar3 is involved in learning and memory in mice', *Frontiers in Neuroscience* **12**(APR), 353124.

Montiel-Gonźalez, M. F., Vallecillo-Viejo, I. C. and Rosenthal, J. J. (2016), 'An efficient system for selectively altering genetic information within mRNAs', *Nucleic Acids Research* **44**(21), e157–e157.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008), 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods 2008 5:7* **5**(7), 621–628.

Müller, J. M., Moos, K., Baar, T., Maier, K. C., Zumer, K. and Tresch, A. (2024), 'Nuclear export is a limiting factor in eukaryotic mRNA metabolism', *PLOS Computational Biology* **20**(5), e1012059.

Musch, M. W., Kapil, A. and Chang, E. B. (2004), 'Heat shock protein 72 binds and protects dihydrofolate reductase against oxidative injury', *Biochemical and Biophysical Research Communications* **313**(1), 185–192.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008), 'The transcriptional landscape of the yeast genome defined by RNA sequencing', *Science* **320**(5881), 1344–1349.

Nagarajan, V. K., Jones, C. I., Newbury, S. F. and Green, P. J. (2013), 'XRN 5ââ3â exoribonucleases: Structure, mechanisms and functions', *Biochimica et biophysica acta* **1829**(0), 590.

Nguyen, T. A., Heng, J. W. J., Kaewsapsak, P., Kok, E. P. L., Stanojević, D., Liu, H., Cardilla, A., Praditya, A., Yi, Z., Lin, M., Aw, J. G. A., Ho, Y. Y., Peh, K. L. E., Wang, Y., Zhong, Q., Heraud-Farlow, J., Xue, S., Reversade, B., Walkley, C., Ho, Y. S., Šikić, M., Wan, Y. and Tan, M. H. (2022), 'Direct identification of A-to-I editing sites with nanopore native RNA sequencing', *Nature Methods 2022 19:7* **19**(7), 833–844.

Nishikura, K. (2015), 'A-to-I editing of coding and non-coding RNAs by ADARs', *Nature Reviews Molecular Cell Biology 2015 17:2* **17**(2), 83–96.

or Karlström, V., do Sagredo, E. A., Planells, J., lotte elinder, C. W., Jungfleisch, J., Barrera-Conde, A., Engfors, L., Daniel, C., Gebauer, F., Visa, N. and Öhman, M.

(2024), 'ADAR3 modulates neuronal differentiation and regulates mRNA stability and translation', *Nucleic Acids Research* **52**(19), 12021–12038.

Ota, S., Tanaka, Y., Yasutake, R., Ikeda, Y., Yuki, R., Nakayama, Y. and Saito, Y. (2023), 'Distinct effects of heat shock temperatures on mitotic progression by influencing the spindle assembly checkpoint', *Experimental cell research* **429**(2).

Park, C., Jin, Y., Kim, Y. J., Jeong, H. and Seong, B. L. (2020), 'RNA-binding as chaperones of DNA binding proteins from starved cells', *Biochemical and biophysical research communications* **524**(2), 484–489.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. (2017), 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods 2017 14:4* **14**(4), 417–419.

Pestal, K., Funk, C. C., Snyder, J. M., Price, N. D., Treuting, P. M. and Stetson, D. B. (2015), 'Isoforms of RNA-Editing Enzyme ADAR1 Independently Control Nucleic Acid Sensor MDA5-Driven Autoimmunity and Multi-organ Development', *Immunity* **43**(5), 933–944.

Philpott, M., Watson, J., Thakurta, A., Brown, T., Brown, T., Oppermann, U. and Cribbs, A. P. (2021), 'Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq', *Nature Biotechnology 2021 39:12* **39**(12), 1517–1520.

Picardi, E., D'Erchia, A. M., Giudice, C. L. and Pesole, G. (2017), 'REDIportal: a comprehensive database of A-to-I RNA editing events in humans', *Nucleic Acids Research* **45**(D1), D750–D757.

Picelli, S., Björklund, Ã. K., Faridani, O. R., Sagasser, S., Winberg, G. and Sandberg, R. (2013), 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nature Methods 2013 10:11* **10**(11), 1096–1098.

Picelli, S., Faridani, O. R., Björklund, Ã. K., Winberg, G., Sagasser, S. and Sandberg, R. (2014), 'Full-length RNA-seq from single cells using Smart-seq2', *Nature Protocols 2013 9:1* **9**(1), 171–181.

Piechotta, M., Wang, Q., Altmüller, J. and Dieterich, C. (2021), 'RNA modification mapping with JACUSA2', *bioRxiv* p. 2021.07.02.450888.

Piechotta, M., Wyler, E., Ohler, U., Landthaler, M. and Dieterich, C. (2017), 'JACUSA: Site-specific identification of RNA editing events from replicate sequencing data', *BMC Bioinformatics* **18**(1), 1–15.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C. and Caracausi, M. (2019), 'Human protein-coding genes and gene feature statistics in 2019', *BMC Research Notes* **12**(1), 1–5.

Poulsen, H., Nilsson, J., Damgaard, C. K., Egebjerg, J. and Kjems, J. (2001), 'CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain', *Molecular and cellular biology* **21**(22), 7862–7871.

Prjibelski, A., Mikheenko, A., Joglekar, A., Smetanin, A., Lapidus, A. and Tilgner, H. (2022), 'IsoQuant: a tool for accurate novel isoform discovery with long reads'.

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. and Regev, A. (2011), 'Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells', *Nature Biotechnology 2011 29:5* **29**(5), 436–442.

Raghava Kurup, R., Oakes, E. K., Manning, A. C., Mukherjee, P., Vadlamani, P. and Hundley, H. A. (2022), 'RNA binding by ADAR3 inhibits adenosine-to-inosine editing and promotes expression of immune response protein MAVS', *Journal of Biological Chemistry* **298**(9), 102267.

Raghava Kurup, R., Oakes, E. K., Vadlamani, P., Nwosu, O., Danthi, P. and Hundley, H. A. (2022), 'ADAR3 activates NF-κB signaling and promotes glioblastoma cell resistance to temozolomide', *Scientific Reports 2022 12:1* **12**(1), 1–15.

Ramaswami, G. and Li, J. B. (2014), 'RADAR: a rigorously annotated database of A-to-I RNA editing', *Nucleic Acids Research* **42**(D1), D109–D113.

Rambout, X. and Maquat, L. E. (2024), 'Nuclear mRNA decay: regulatory networks that control gene expression', *Nature Reviews Genetics 2024 25:10* **25**(10), 679–697.

Ramsköld, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P. and Sandberg, R. (2012), 'Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells', *Nature Biotechnology 2012 30:8* **30**(8), 777–782.

Ravi, V. M., Joseph, K., Wurm, J., Behringer, S., Garrelfs, N., dâErrico, P., Naseri, Y., Franco, P., Meyer-Luehmann, M., Sankowski, R., Shah, M. J., Mader, I., Delev, D., Follo, M., Beck, J., Schnell, O., Hofmann, U. G. and Heiland, D. H. (2019), 'Human organotypic brain slice culture: A novel framework for environmental research in neuro-oncology', *Life Science Alliance* **2**(4).

Rice, G. I., Kasher, P. R., Forte, G. M., Mannion, N. M., Greenwood, S. M., Szynkiewicz, M., Dickerson, J. E., Bhaskar, S. S., Zampini, M., Briggs, T. A., Jenkinson, E. M., Bacino, C. A., Battini, R., Bertini, E., Brogan, P. A., Brueton, L. A., Carpanelli, M., De Laet, C., De Lonlay, P., Del Toro, M., Desguerre, I., Fazzi, E., Garcia-Cazorla, Ã., Heiberg, A., Kawaguchi, M., Kumar, R., Lin, J. P. S., Lourenco, C. M., Male, A. M., Marques, W., Mignot, C., Olivieri, I., Orcesi, S., Prabhakar, P., Rasmussen, M., Robinson, R. A., Rozenberg, F., Schmidt, J. L., Steindl, K., Tan, T. Y., Van Der Merwe, W. G., Vanderver, A., Vassallo, G., Wakeling, E. L., Wassmer, E., Whittaker, E., Livingston, J. H., Lebon, P., Suzuki, T., McLaughlin, P. J., Keegan, L. P., O'Connell, M. A., Lovell, S. C. and Crow, Y. J. (2012), 'Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature', *Nature Genetics 2012 44:11* **44**(11), 1243–1248.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. (2011), 'Integrative Genomics Viewer', *Nature biotechnology* **29**(1), 24.

Rodrigues, D. C., Mufteev, M., Yuki, K. E., Narula, A., Wei, W., Piekna, A., Liu, J., Pasceri, P., Rissland, O. S., Wilson, M. D. and Ellis, J. (2023), 'Buffering

of transcription rate by mRNA half-life is a conserved feature of Rett syndrome models', *Nature Communications 2023 14:1* **14**(1), 1–15.

Rodriques, S. G., Chen, L. M., Liu, S., Zhong, E. D., Scherrer, J. R., Boyden, E. S. and Chen, F. (2020), 'RNA timestamps identify the age of single molecules in RNA sequencing', *Nature Biotechnology 2020 39:3* **39**(3), 320–325.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F. and Macosko, E. Z. (2019), 'Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution', *Science* **363**(6434), 1463–1467.

Roundtree, I. A., Evans, M. E., Pan, T. and He, C. (2017), 'Dynamic RNA Modifications in Gene Expression Regulation', *Cell* **169**(7), 1187–1200.

Rummel, T., Sakellaridi, L. and Erhard, F. (2023), 'grandR: a comprehensive package for nucleotide conversion RNA-seq data analysis', *Nature Communications 2023 14:1* **14**(1), 1–17.

Salmen, F., De Jonghe, J., Kaminski, T. S., Alemany, A., Parada, G. E., Verity-Legg, J., Yanagida, A., Kohler, T. N., Battich, N., van den Brekel, F., Ellermann, A. L., Arias, A. M., Nichols, J., Hemberg, M., Hollfelder, F. and van Oudenaarden, A. (2022), 'High-throughput total RNA sequencing in single cells using VASA-seq', *Nature Biotechnology 2022 40:12* **40**(12), 1780–1793.

Salpietro, V., Dixon, C. L., Guo, H., Bello, O. D., Vandrovcova, J., Efthymiou, S., Maroofian, R., Heimer, G., Burglen, L., Valence, S., Torti, E., Hacke, M., Rankin, J., Tariq, H., Colin, E., Procaccio, V., Striano, P., Mankad, K., Lieb, A., Chen, S., Pisani, L., Bettencourt, C., Männikkö, R., Manole, A., Brusco, A., Grosso, E., Ferrero, G. B., Armstrong-Moron, J., Gueden, S., Bar-Yosef, O., Tzadok, M., Monaghan, K. G., Santiago-Sim, T., Person, R. E., Cho, M. T., Willaert, R., Yoo, Y., Chae, J. H., Quan, Y., Wu, H., Wang, T., Bernier, R. A., Xia, K., Blesson, A., Jain, M., Motazacker, M. M., Jaeger, B., Schneider, A. L., Boysen, K., Muir, A. M., Myers, C. T., Gavrilova, R. H., Gunderson, L., Schultz-Rogers, L., Klee, E. W., Dyment, D., Osmond, M., Parellada, M., Llorente, C., Gonzalez-Peñas, J., Carracedo, A., Van Haeringen, A., Ruivenkamp, C., Nava, C., Heron, D., Nardello, R., Iacomino, M., Minetti, C., Skabar, A., Fabretto, A., Hanna, M. G., Bugiardini, E., Hostettler, I., OâCallaghan, B., Khan, A., Cortese, A., OâConnor, E., Yau, W. Y., Bourinaris, T., Kaiyrzhanov, R., Chelban, V., Madej, M., Diana, M. C., Vari, M. S., Pedemonte, M., Bruno, C., Balagura, G., Scala, M., Fiorillo, C., Nobili, L., Malintan, N. T., Zanetti, M. N., Krishnakumar, S. S., Lignani, G., Jepson, J. E., Broda, P., Baldassari, S., Rossi, P., Fruscione, F., Madia, F., Traverso, M., De-Marco, P., Pérez-Dueñas, B., Munell, F., Kriouile, Y., El-Khorassani, M., Karashova, B., Avdjieva, D., Kathom, H., Tincheva, R., Van-Maldergem, L., Nachbauer, W., Boesch, S., Gagliano, A., Amadori, E., Goraya, J. S., Sultan, T., Kirmani, S., Ibrahim, S., Jan, F., Mine, J., Banu, S., Veggiotti, P., Zuccotti, G. V., Ferrari, M. D., Van Den Maagdenberg, A. M., Verrotti, A., Marseglia, G. L., Savasta, S., Soler, M. A., Scuderi, C., Borgione, E., Chimenz, R., Gitto, E., Dipasquale, V., Sallemi, A., Fusco, M., Cuppari, C., Cutrupi, M. C., Ruggieri, M., Cama, A., Capra, V., Mencacci, N. E., Boles, R., Gupta, N., Kabra, M., Papacostas, S., Zamba-Papanicolaou, E., Dardiotis, E., Maqbool, S., Rana,

N., Atawneh, O., Lim, S. Y., Shaikh, F., Koutsis, G., Breza, M., Coviello, D. A., Dauvilliers, Y. A., AlKhawaja, I., AlKhawaja, M., Al-Mutairi, F., Stojkovic, T., Ferrucci, V., Zollo, M., Alkuraya, F. S., Kinali, M., Sherifa, H., Benrhouma, H., Turki, I. B., Tazir, M., Obeid, M., Bakhtadze, S., Saadi, N. W., Zaki, M. S., Triki, C. C., Benfenati, F., Gustincich, S., Kara, M., Belcastro, V., Specchio, N., Capovilla, G., Karimiani, E. G., Salih, A. M., Okubadejo, N. U., Ojo, O. O., Oshinaike, O. O., Oguntunde, O., Wahab, K., Bello, A. H., Abubakar, S., Obiabo, Y., Nwazor, E., Ekenze, O., Williams, U., Iyagba, A., Taiwo, L., Komolafe, M., Senkevich, K., Shashkin, C., Zharkynbekova, N., Koneyev, K., Manizha, G., Isrofilov, M., Guliyeva, U., Salayev, K., Khachatryan, S., Rossi, S., Silvestri, G., Haridy, N., Ramenghi, L. A., Xiromerisiou, G., David, E., Aguennouz, M., Fidani, L., Spanaki, C., Tucci, A., Raspall-Chaure, M., Chez, M., Tsai, A., Fassi, E., Shinawi, M., Constantino, J. N., De Zorzi, R., Fortuna, S., Kok, F., Keren, B., Bonneau, D., Choi, M., Benzeev, B., Zara, F., Mefford, H. C., Scheffer, I. E., Clayton-Smith, J., Macaya, A., Rothman, J. E., Eichler, E. E., Kullmann, D. M. and Houlden, H. (2019), 'AMPA receptor GluA2 subunit defects are a cause of neurodevelopmental disorders', *Nature Communications 2019 10:1* **10**(1), 1–16.

Sanger, F. and Coulson, A. R. (1975), 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', *Journal of Molecular Biology* **94**(3), 441–448.

Sato, T., Stange, D. E., Ferrante, M., Vries, R. G., Van Es, J. H., Van Den Brink, S., Van Houdt, W. J., Pronk, A., Van Gorp, J., Siersema, P. D. and Clevers, H. (2011), 'Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium', *Gastroenterology* **141**(5), 1762–1772.

Schaffer, A. A., Kopel, E., Hendel, A., Picardi, E., Levanon, E. Y. and Eisenberg, E. (2020), 'The cell line A-to-I RNA editing catalogue', *Nucleic Acids Research* **48**(11), 5849–5858.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995), 'Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray', *Science* **270**(5235), 467–470.

Schoenberg, D. R. and Maquat, L. E. (2012), 'Regulation of cytoplasmic mRNA decay', *Nature Reviews Genetics 2012 13:4* **13**(4), 246–259.

Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. and Simon, M. D. (2018), 'TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding', *Nature Methods 2018 15:3* **15**(3), 221–225.

Schueler, M., Munschauer, M., Gregersen, L. H., Finzel, A., Loewer, A., Chen, W., Landthaler, M. and Dieterich, C. (2014), 'Differential protein occupancy profiling of the mRNA transcriptome', *Genome biology* **15**(1).

Shi, C. and Pamer, E. G. (2011), 'Monocyte recruitment during infection and inflammation', *Nature Reviews Immunology 2011 11:11* **11**(11), 762–774.

Shi, Y., Kirwan, P. and Livesey, F. J. (2012), 'Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks', *Nature Protocols 2012 7:10* **7**(10), 1836–1846.

Stefl, R., Xu, M., Skrisovska, L., Emeson, R. B. and Allain, F. H. (2006), 'Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs', *Structure (London, England : 1993)* **14**(2), 345–355.

Sun, J., Philpott, M., Loi, D., Li, S., Monteagudo-Mesas, P., Hoffman, G., Robson, J., Mehta, N., Gamble, V., Brown, T., Brown, T., Canzar, S., Oppermann, U. and Cribbs, A. P. (2024), 'Correcting PCR amplification errors in unique molecular identifiers to generate accurate numbers of sequencing molecules', *Nature Methods 2024 21:3* **21**(3), 401–405.

Sun, T., Yu, Y., Wu, X., Acevedo, A., Luo, J. D., Wang, J., Schneider, W. M., Hurwitz, B., Rosenberg, B. R., Chung, H. and Rice, C. M. (2021), 'Decoupling expression and editing preferences of ADAR1 p150 and p110 isoforms', *Proceedings of the National Academy of Sciences of the United States of America* **118**(12), e2021757118.

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., Liu, K. I., Zhang, R., Ramaswami, G., Ariyoshi, K., Gupte, A., Keegan, L. P., George, C. X., Ramu, A., Huang, N., Pollina, E. A., Leeman, D. S., Rustighi, A., Goh, Y. P., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Karczewski, K. J., Lek, M., Li, X., MacArthur, D. G., Nedzel, J. L., Nguyen, D. T., Noble, M. S., Segrè, A. V., Trowbridge, C. A., Tukiainen, T., Abell, N. S., Balliu, B., Barshir, R., Basha, O., Battle, A., Bogu, G. K., Brown, A., Brown, C. D., Castel, S. E., Chen, L. S., Chiang, C., Conrad, D. F., Cox, N. J., Damani, F. N., Davis, J. R., Delaneau, O., Dermitzakis, E. T., Engelhardt, B. E., Eskin, E., Ferreira, P. G., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D., Gliner, G., Gloudemans, M. J., Guigo, R., Hall, I. M., Han, B., He, Y., Hormozdiari, F., Howald, C., Im, H. K., Jo, B., Kang, E. Y., Kim, Y., Kim-Hellmuth, S., Lappalainen, T., Li, G., Li, X., Liu, B., Mangul, S., McCarthy, M. I., McDowell, I. C., Mohammadi, P., Monlong, J., Montgomery, S. B., Muñoz-Aguirre, M., Ndungu, A. W., Nicolae, D. L., Nobel, A. B., Oliva, M., Ongen, H., Palowitch, J. J., Panousis, N., Papasaikas, P., Park, Y., Parsana, P., Payne, A. J., Peterson, C. B., Quan, J., Reverter, F., Sabatti, C., Saha, A., Sammeth, M., Scott, A. J., Shabalin, A. A., Sodaei, R., Stephens, M., Stranger, B. E., Strober, B. J., Sul, J. H., Tsang, E. K., Urbut, S., Van De Bunt, M., Wang, G., Wen, X., Wright, F. A., Xi, H. S., Yeger-Lotem, E., Zappala, Z., Zaugg, J. B., Zhou, Y. H., Akey, J. M., Bates, D., Chan, J., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J. A., Feinberg, A. P., Fernando, M. S., Halow, J., Hansen, K. D., Haugen, E., Hickey, P. F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kellis, M., Kibriya, M. G., Lee, K., Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M. T., Molinie, B., Nelson, J., Neri, F. J., Park, Y., Pierce, B. L., Rinaldi, N. J., Rizzardi, L. F., Sandstrom, R., Skol, A., Smith, K. S., Snyder, M. P., Stamatoyannopoulos, J., Tang, H., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Nierras, C. R., Branton, P. A., Carithers, L. J., Guan, P., Moore, H. M., Rao, A., Vaught, J. B., Gould, S. E., Lockart, N. C., Martin, C., Struewing, J. P., Volpi, S., Addington, A. M., Koester, S. E., Little, A. R., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar,

R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D. R., Davis, D. A., Mash, D. C., Undale, A. H., Smith, A. M., Tabor, D. E., Roche, N. V., McLean, J. A., Vatanian, N., Robinson, K. L., Sobin, L., Barcus, M. E., Valentino, K. M., Qi, L., Hunter, S., Hariharan, P., Singh, S., Um, K. S., Matose, T., Tomaszewski, M. M., Barker, L. K., Mosavel, M., Siminoff, L. A., Traino, H. M., Flicek, P., Juettemann, T., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S. J., Zerbino, D. R., Craft, B., Goldman, M., Haeussler, M., Kent, W. J., Lee, C. M., Paten, B., Rosenbloom, K. R., Vivian, J., Zhu, J., Chawla, A., Del Sal, G., Peltz, G., Brunet, A., Samuel, C. E., O'Connell, M. A., Walkley, C. R., Nishikura, K. and Li, J. B. (2017), 'Dynamic landscape and regulation of RNA editing in mammals', *Nature 2017 550:7675* **550**(7675), 249–254.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. and Surani, M. A. (2009), 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nature Methods 2009 6:5* **6**(5), 377–382.

Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y. and Akimitsu, N. (2012), 'Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals', *Genome research* **22**(5), 947–956.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J. R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A. C., Johannessen, C. M., Andreev, A. Y., Van Allen, E. M., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jané-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A. and Garraway, L. A. (2016), 'Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq', *Science* **352**(6282), 189–196.

Ura, H., Togi, S. and Niida, Y. (2022), 'A comparison of mRNA sequencing (RNA-Seq) library preparation methods for transcriptome analysis', *BMC Genomics* **23**(1), 1–10.

Viegas, J. O., Fishman, L., Meshorer, E. and Rabani, M. (2023), 'Calculating RNA degradation rates using large-scale normalization in mouse embryonic stem cells', *STAR Protocols* **4**(3), 102534.

Vogel, P., Moschref, M., Li, Q., Merkle, T., Selvasaravanan, K. D., Li, J. B. and Stafforst, T. (2018), 'Efficient and precise editing of endogenous transcripts with SNAP-tagged ADARs', *Nature methods* **15**(7), 535–538.

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T. and He, C. (2014), 'N6-methyladenosine-dependent regulation of messenger RNA stability', *Nature* **505**(7481), 117–120.

Wang, Z., Gerstein, M. and Snyder, M. (2009), 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature reviews. Genetics* **10**(1), 57.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A.,

Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R. and Hunkapiller, M. W. (2019), 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature Biotechnology 2019 37:10* **37**(10), 1155–1162.

Wolf, J., Gerber, A. P. and Keller, W. (2002), 'tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli', *The EMBO journal* **21**(14), 3841–3851.

Xiang, J. F., Yang, Q., Liu, C. X., Wu, M., Chen, L. L. and Yang, L. (2018), 'N6-Methyladenosines Modulate A-to-I RNA Editing', *Molecular Cell* **69**(1), 126–135.

Xiao, Y. L., Liu, S., Ge, R., Wu, Y., He, C., Chen, M. and Tang, W. (2023), 'Transcriptome-wide profiling and quantification of N6-methyladenosine by enzyme-assisted adenosine deamination', *Nature biotechnology* **41**(7), 993–1003.

Yaish, O. and Orenstein, Y. (2022), 'Computational modeling of mRNA degradation dynamics using deep neural networks', *Bioinformatics (Oxford, England)* **38**(4), 1087–1101.

Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M. and Darnell, J. E. (2003), 'Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes', *Genome Research* **13**(8), 1863.

Yi, H., Park, J., Ha, M., Lim, J., Chang, H. and Kim, V. N. (2018), 'PABP Cooperates with the CCR4-NOT Complex to Promote mRNA Deadenylation and Block Precocious Decay', *Molecular Cell* **70**(6), 1081–1088.

Zhang, X., Devany, E., Murphy, M. R., Glazman, G., Persaud, M. and Kleiman, F. E. (2015), 'PARN deadenylase is involved in miRNA-dependent degradation of TP53 mRNA in mammalian cells', *Nucleic Acids Research* **43**(22), 10925–10938.

Zheng, S. C., Stein-OâBrien, G., Boukas, L., Goff, L. A. and Hansen, K. D. (2023), 'Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates', *Genome Biology* **24**(1), 1–31.