THE UNIVERSITY OF CHICAGO

DECIPHERING THE ROLE OF RNA PROCESSING IN HUMAN BIOLOGY AND DISEASE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY
ANKEETA SHAH

CHICAGO, ILLINOIS
JUNE 2023

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I am grateful for all of the guidance, support, and mentorship I received during my PhD. In particular, none of my projects would have been possible without my PhD advisor, Yang Li, who helped me develop as a scientist, thinker, and programmer. I was extremely fortunate that he took me into his lab as his first graduate student, and I will take the skills that I have developed and appreciation of scientific research in whatever I pursue in my career.

I am indebted to the current and past members of the Li lab. In particular, I would like to thank Fabio Morgante, Yi Zeng, Benjamin Fair, and Phoenix Mu for their incisive questions and input on various projects. I am thankful for my committee members, Xiaochang Zhang, Jon Staley, and Alex Ruthenburg for insightful discussions. I have learned a tremendous amount about human genetics, RNA biology, and neuroscience from all of you. Many of my projects would not have been possible without the help of collaborators in the Gilad lab at the University of Chicago, the Popko lab at Northwestern (formerly at the University of Chicago), and the Eichler lab at the University of Washington.

I was very fortunate to have the help of Sue Levison, our dedicated administrator, who has supported the Committee on Genetics, Genomics, and Systems Biology in inumerable of ways. Furthermore, I was fortunate to have been funded by the Genetics and Regulation Training Grant (T32GM007197) from the National Institutes of Health from 2017-2020, which was a grant that Lucia Rothman-Denes managed.

Thank you to the friends that I made at the University of Chicago. It was an honor to learn about science from you, share coffees and meals, and discuss life. In particular, Maryn Carlson, the polymath, thank you for your friendship over the years. You have and will continue to inspire me. Thank you to the friends that I made through the Amgen Scholars Program in 2015, namely Brian Ho and Yuzhang Chen, who never failed to pick up my phone calls over the years. You are brilliant scientists, and I look forward to seeing the strides you make in your respective fields. Thank you to my non-scientist friends in Chicago and New

York for your patience, visits, and unwavering support.

Several mentors were instrumental in inspiring me to pursue science and a PhD. Thank you to Chaolin Zhang at Columbia University for taking me into your lab when I knew nothing about bioinformatics when I was a naive but optimistic undergraduate student. You were the first person to get me enthusiastic about RNA biology and the prospect of commercializing scientific innovations. And thank you to Tom Callahan, formerly my high school science research teacher at Dobbs Ferry High School, for instilling in me a love of research, critical thinking, and creativity.

Lastly, thank you to my family for allowing me to pursue my passions with no judgement. My mother, father, and grandmother courageously emigrated to the United States from India in the 1980s with nothing. They worked a myriad of odd jobs to stay afloat, and they taught me a tremendous amount about hard work and resilience. This PhD is as much mine as it is theirs.

ABSTRACT

Genome-wide association studies (GWAS) have allowed us to successfully identify thousands of common genetic variants underlying a number of diseases, but it has been difficult to understand how mechanism of action because the vast majority of these loci are located in non-coding regions of the genome. Because it is estimated that only 25% of disease associated genetic variants contribute to disease by affecting steady-state gene expression levels, it will be important to establish a more comprehensive understanding of alternative mechanisms through which genetic variants act to contribute to disease, such as RNA processing and RNA modification. Motivated by this, this dissertation outlines the development of computational methods and assessment of existing tools to profiles various RNA processing and RNA modification events across individuals, cell types, and developmental stages, which can ultimately be applied to large disease-cohort datasets in future studies. In the first chapter, we provide primers on quantitative genetics and RNA processing and modifications to put this work in context. In the second chapter, we demonstrate that combining large quantities of RNA-seq data with small quantities of specialized data, including 3'-Seq and single-molecule real-time (SMRT) isoform sequencing (Iso-Seq), allows one to study alternative cleavage and polyadenylation, without compromising affordability or accuracy. We apply this approach to explore inter-individual variation in polyadenylation site choice. In the third chapter, we profile the role of the RNA modification N6-methyladenosine (m6A) in oligodendrocyte lineage progression and its potential impacts in human diseases, such as multiple sclerosis. Finally, in the fourth chapter, we develop a method to examine how genetic variants that increase risk of disease reduce the fidelity of RNA splicing.

CHAPTER 1

INTRODUCTION

1.1 A primer on quantitative genetics

Quantitative traits, such as height [93], or molecular phenotypes, such as gene expression, vary continuously, and often follow a normal distribution. In fact, most observable or phenotypic variation between individuals in a population tends to be quantitative. The field of quantitative genetics aims to understand the genetic factors that contribute to phenotypic variation in a population. The process of identifying genetic loci that are associated with phenotypic variation is known as quantitative trait loci (QTL) mapping [93]. Over the last twenty years, QTL mapping has been used extensively in a wide variety of biological contexts, including agriculture, medical genetics, evolution, and functional genomics [93].

Unlike linkage-baed QTL mapping [93], which leans on small pedigree data to allow researchers to study the transmission of a particular locus within a family, association mapping of a sample of individuals from a large population is now routine, particularly at a genome-wide scale (i.e. genome-wide association studies (GWAS). The intuition behind this shift in approach is that, at the population level, a larger number of recombination events are able to reduce linkage disequilibrium (LD), which would normally restrict the resolution of QTL mapping of individual loci. Therefore, association studies allow for much finer resolution mapping of QTL. In fact, within the last decade, GWAS have successfully identified thousands of common genetic variants, or single nucleotide polymorphisms (SNPs) [2, 22], associated with complex traits and diseases [130, 1, 189, 80].

One limitation of GWAS is that they do not reveal which genes or mechanisms may causally link these common genetic variants to a trait or disease. Moreover, elucidating the mechanisms underlying complex traits or diseases requires detailed understanding of how loci implicated in GWAS impact the disease in the relevant cell type(s), which is often difficult

to discern. Because it is estimated that over 90% of disease-associated variants identified by GWAS are in non-coding regions of the genome [114], it is likely that these genetic variants act through *gene regulatory* mechanisms.

1.1.1 Understanding disease pathogenesis by integrating GWAS hits with functional data

One successful approach to investigate the mechanisms by which noncoding genetic variants act to contribute to phenotypic variation is *molecular* QTL mapping, whereby one is able to identify associations between SNPs and molecular phenotypes, such as gene expression, chromatin accessibility, and protein levels. In 2010, through intersections of GWAS and molecular QTL data, Nicolae and colleagues identified strong enrichment of common genetic variants that affect gene expression levels (i.e. expression QTL (eQTL)) [126]. This enrichment implies that SNPs that contribute to variation in phenotypic outcomes do so by either increasing or decreasing gene expression.

This motivated a number of follow-up studies that identified and integrated eQTL with disease associated GWAS hits [136, 94] across different tissues, cell-types, and contexts, in the hope of improving our understanding of how trait-associated variants mechanistically act to affect disease. Such work has successfully identified a subset of SNPs that directly disrupt cis-regulatory elements, such as promoters and enhancers, for example, by disrupting transcription factor binding [94, 119, 20, 37] and chromatin accessibility (as measured by DNase I sensitivity or H3K27ac) [14, 40, 118], contributing to gene expression level dysregulation.

1.1.2 A putative mechanism underlying disease pathogenesis: disruption of RNA processing events

In 2016, Li and colleagues systematically explored the effects of genetic variation across all stages of gene regulation, from chromatin to protein, in the HapMap/1000 Genomes Yoruba collection of lymphoblastoid cell lines (LCLs) [104, 34, 118]. Interestingly, the authors determined that there are several mechanisms, in addition to gene expression regulation at the promoter and enhancer level, that play important roles in linking genetic variation to complex traits and disease risk. Moreover, a study published the year after estimated that up to 75% of the time the causal variant that underlies variation in a particular trait is distinct from the causal variant that underlies variation in gene expression [29]. This implies that the majority of these disease-associated variants may act independently of promoters and enhancers, functioning through a gene regulatory mechanism distinct from gene expression.

To obtain models of how genetic variation impacts complex traits and disease and to predict the functional impact of non-coding genetic variants, we need a better understanding of regulatory mechanisms acting at every stage of the gene regulatory cascade, not just gene expression. It is plausible that a substantial number of single nucleotide polymorphisms (SNPs) that do not contribute to variation in gene expression affect disease risk by disrupting the regulation of various RNA processing events, such as alternative splicing or alternative polyadenylation, resulting in shifts in the relative proportions of mRNA isoforms produced. By identifying common genetic variants that affect alternative splicing or alternative polyadenylation by mapping QTL for these molecular phenotypes, one might be able to better understand the mechanisms through which these SNPs act, such as disrupting of splice sites or polyadenylation signal sites, to contribute to variation in disease outcomes.

1.2 A primer on RNA processing events

A central question in genetics is how diversity in the human transcriptome is achieved despite the human genome only encoding 20,000 protein coding genes [45]. Over the last two decades, it has become increasingly apparent that the astonishing complexity within the human transcriptome can be attributed to RNA processing events, such as the use of alternative transcription start sites, alternative splicing, and alternative polyadenylation, allowing for a single gene to encode a repertoire of at least 200,000 unique transcripts [45]. RNA processing events, which allow for the transcriptome to vary dynamically across different tissues, developmental stages, and disease states, have been recent targets of systematic study, which has been aided tremendously by one of the most significant technological breakthroughs in genomics, the development of RNA-sequencing (RNA-seq) [91]. RNA-seq has allowed researchers to, without a priori knowledge of gene annotations, quantitatively measure gene expression, discover novel transcripts, and measure the relative abundance of distinct transcript isoforms.

1.2.1 Alternative polyadenylation

Alternative polyadenylation (APA), or the process by which a single gene is able to produce multiple mRNA isoforms with distinct 3' ends (e.g. mRNA species of different 3' untranslated regions (UTRs) is a critical RNA processing event that allows a single gene to encode multiple mRNA transcripts. In fact, APA also can affect the stability, localization, transport, and translation of mRNA [168]. APA involves a four component multi-protein complex, including CPSF (Cleavage and Polyadenylation Specificity Factor), CstF (Cleavage stimulation Factor), CFI and II (Cleavage Factors I and II), poly(A) polymerase (PAP), and accessary factors [141]. The assembly of this 3' end processing complex on pre-mRNA involves the interaction of CPSF with the degenerate polyadenylation signal site (PAS), AAUAAA, which is located 20-30 nucleotides upstream of the cleavage site, and CstF with

the downstream U/GU-rich sequence [141]. This results in cleavage and subsequent addition of a polyA tail that can vary in length [166].

Usage of a particular poly(A) site over another is determined by the relative strength of the PAS, similar to usage of splice sites based on their strength. Moreover, the auxiliary factors involved in PAS choice are analogous to RNA binding proteins that also help mediate splicing choice. For example, knockdown of the termination factors Pcf11 and Fip1 contributes to increased usage of proximal sites over distal sites in a wide range of genes [101]. By using different PASs, genes can either shorten or extend, for example, their 3'UTRs, which allows a transcript to contain distinct cis-regulatory elements, such as miRNA binding sites or RNA-binding protein sites [115], which can be important in the regulation of normal development [8, 162], or mis-regulation in disease [95].

Early studies mapping APA were done on single genes; one motivating example is the gene encoding the immunoglobulin M heavy chain (IgM H chain) [8, 162, 133]. The first antibodies produced by nave or memory B cells are not secreted, and instead are inserted into the plasma membrane, where they serve as antigen receptors. However, upon B cell activation by an antigen, the B cell proliferates and differentiates into an antibody secreting effector cell, namely, a plasma cell, which produces soluble, or secreted antibodies. However, for a long time, immunologists did not have a deep, mechanistic understanding of what drove this shift in antibody localization, despite the cells still encoding antibodies specific for the same antigen. It was later elucidated that intronic (proximal) versus 3'UTR (distal) PAS usage in the antibody-encoding gene, IgM H, during B cell differentiation is what drives antibody localization or solubility, and specifically this is known to be mediated by the differences in the concentration of the CstF. CstF concentration is low in undifferentiated B cells such that it will preferentially bind the stronger GU sequence adjacent to the distal PAS, allowing for splicing out of the intervening intronic sequences and resulting in a longer, membrane bound IgM. In contrast, upon differentiation, in plasma cells, the concentration

of CstF is high such that CstF is able to bind the weaker GU sequence at the proximal PAS, not allowing for splicing out of said intron, resulting in a shorter, secreted form of IgM [162]. This example highlights the critical importance of APA in mediating the production of the correct isoform, membrane bound or secreted, of IgM in the correct cell type along a particular differentiation trajectory.

Recent efforts have been made in the context of cancer biology and immunobiology to better understand PAS choice or usage in distinct cellular contexts more broadly across the human genome [141, 117, 158]. Specifically, RNA-seq studies have revealed that at least 70% of genes in the human genome have multiple polyadenylation sites, some of which are in introns, and most of which are in the 3' UTR [42]. Because APA, like alternative splicing, seems to be a pervasive RNA processing phenomenon that is critical in cell-type specific regulation, an important next step will be to characterize how SNPs can impact PAS choice, and how this can lead to variation in phenotypic outcomes. For example, Cannovo et al. studied the effect of genetic variation on post-transcriptional 3' RNA processing regulation across multiple stages of metazoan development in wild Drosophila isolates [23]. The authors discovered thousands of alternative polyadenylation QTL (apaQTL), which were enriched outside of enhancers, highlighting the importance of studying gene regulatory mechanisms beyond genetic elements that impact gene expression levels. To date, very few studies have quantified APA in human population samples and detected variants implicated in genome-wide APA variation [99].

1.2.2 RNA methylation

Epigenetic regulation spans both DNA and RNA [56]. The recent discovery of reversible mRNA methylation has revealed a new dimension of post-transcriptional gene regulation [192]. m6A is a chemical derivative of adenosine (A) in RNA, and it is the most abundant RNA modification observed in mammals across both polyadenylated mRNA and non-coding

RNA, occurring at a frequency of 1-2% per nucleotide [85]. However, due to limitations associated with techniques currently utilized to study m6A the number of m6A marked transcripts is likely underestimated [47].

The m6A modification plays a critical role in a myriad of regulatory processes, including mRNA export [148], mRNA stability[177], and translation [178][202][96]. Levels of m6A are dynamically regulated by both writers, namely m6A methyltransferase complexes like the METTL3-METTL14 complex [109], and erasers, namely m6A demethylases as ALKBH5 [201] and FTO [81]. The m6A reader proteins m6A-modified RNAs. Some m6A readers proteins contain the YT521-B homology (YTH) domain while others include several of the heterogeneous nuclear ribonucleoproteins (HNRNPs) [184] [187]. RNA metabolism, including stability, translation, localization, and splicing, can be regulated by m6A [147]. Despite the appreciation of m6A's role in health and disease, very few studies have been conducted to understand the impact of variation in RNA editing on complex genetic diseases. In a recent study, Zhang et al. discovered that m6A QTLs are enriched for risk variants of a range of complex traits, particularly autoimmune diseases and blood cell-related traits [197].

1.2.3 Alternative splicing

The vast majority of human genes contain multiple short exons and long, intervening introns that must be removed from the nascent transcript during mRNA maturation. The exons are joined to form a mRNA that can be translated into a functional protein. The excision of introns from pre-mRNA and the joining of exons is directed by sequences at intron-exon boundaries called splice sites (SS). The GU dinucleotide within the consensus sequence at the 5' end of an intron marks the 5' SS. Near the other end of the intron, there is a branch point, a polypyrimidine tract, and a terminal AG that marks the 3' SS [112]. With this, splicing is carried out by the multiprotein complex called the spliceosome that catalyzes the two transesterification steps of splicing. Specifically, in the first transesterification step, the

2'-hydroxyl of the A nucleotide at the branch point attacks the phosphate at the 5' SS, resulting in cleavage of the 5' exon from the intron and ligation of the intron 5' end with the branch point (e.g. lariat) [112]. The second transesterification step involves the exposed 3'-hydroxyl of the detached exon attacking the phosphate at the 3' end of the intron, allowing for ligation of the two exons and the release of the intron in lariat form [112].

Alternative splicing of pre-mRNA is the process by which use of alternative 5' splice sites, alternative 3' splice sites, casette-exon inclusion or skipping, and intron retention result in the production of distinct, mature transcript isoforms [127, 87]. In fact, it is estimated that at least 95% of human multi-exon genes undergo alternative splicing [174], which is important for diversifying the transcriptome and proteome. The fidelity of splicing is achieved by combinatorial recognition of specific sequences by protein factors within precursor mRNA at many steps during the splicing process [152]. Because about one-third of the human genome is comprised of introns [82], which is a large sequence space containing many sequence elements similar to consensus motifs of canonical splice sites, or cryptic splice sites, it is important that tight quality control mechanisms are in place to prevent the production of a large number of aberrant transcripts. Thus, splicing fidelity is thought to be achieved by a combination of kinetic and thermodynamic mechanisms, including kinetic proofreading which the spliceosome actively rejects suboptimal substrates or sequence elements through ATP-dependent mechanisms mediated by DEAD/H-box ATPases [152]. The thermodynamic mechanism is involved in the catalytic stages of splicing in which the spliceosome preferentially sequesters suboptimal substrates into non-productive conformations that are in equilibrium with the catalytic conformations of the spliceosome [159]. This ultimately prevents the extensive use of suboptimal, or cryptic splice sites, such that it is estimated that the splicing error rate, per intron, is approximately 0.7% [137]. These low-abundance, non-functional transcripts, which tend not to be conserved over evolutionary time, are produced by a process known as noisy splicing, or error-prone use of cryptic splice sites [137].

Li et al. recently showed the critical importance of splicing as a link between genetic variation and disease [104]. Genetic variants can impact RNA splicing by disrupting sequence or directly impacting recognition of canonical splice sites or splicing regulatory elements [195], which can result in aberrant mRNA transcripts, which can cause a large array of human diseases [129, 107]. For example, patients with monotonic dystrophy type 1 have an expanded CUG repeat in the 3'UTR of the DM protein kinase gene, which results in the sequestration of splicing regulators in the muscle-blind protein family, disrupting a number of muscle-blind-dependent splicing events [83, 151]. Thus, mapping common genetic variants that affect RNA splicing (i.e. mapping splicing QTL (sQTL), like mapping eQTL, may lead us to improve the functional interpretation of non-coding disease variants [103]. In fact, several studies have underscored this point by identifying sQTL in high linkage disequilibrium with GWAS associations for a number of different diseases, including Type 2 diabetes, Alzheimer's disease, and schizophrenia [51, 143, 163].

1.3 Dissertation overview

Gene regulation includes a wide range of mechanisms that together allow proper expression of RNA and proteins in a cell-type and developmental-specific manner. Thus far, genetic studies of disease associated variants have primarily focused on effects on steady-state gene expression levels, but these, on average, only account for 25% of disease single nucleotide polymorphisms (SNPs) [29]. Therefore, in this dissertation, I will present my work to study alternative mechanisms, with a focus on RNA processing. **Chapter 1** presents a primer on quantitive genetics and RNA processing and modification. **Chapter 2** evaluates various methods and tools that facilitate the study of alternative polyadenylation. We apply a combination of approaches to explore inter-individual variation in polyadenylation site choice. The work present in this chapter also appears in the journal article Ankeeta Shah, Briana E Mittleman, Yoav Gilad, Yang I Li. Benchmarking sequencing methods and tools that

facilitate the study of alternative polyadenylation. Benchmarking computational tools and methods that facilitate the study of alternative polyadenylation. *Genome Biology* 22 (1), 1-21, (2021). **Chapter 3** investigates the potential role of the m6A mark in regulating differential splicing during oligodendrocyte development. Some of this work also appears in the journal article Huan Xu, Yulia Dzhashiashvili, Ankeeta Shah, Rejani B. Kunjamma, Yi-lan Weng, Benayahu Elbaz, Qili Fei, Joshua S. Jones, Yang I. Li, Xiaoxi Zhuang, Guo-li Ming, Chuan He, and Brian Popko. m6A mRNA Methylation Is Essential for Oligodendrocyte Maturation and CNS Myelination. *Neuron.* 105 (2), 293-309. e5 (2020). **Chapter 4** presents a method to study splicing fidelity and its impact on complex traits and diseases.

In the **appendix**, I have attached abstracts of one book chapter, of which I am the first author, and one additional publications, of which I am a co-author. The book chapter outlines recent progress made and methods used to discover putative regulatory regions associated with complex traits, with a focus on mapping splicing quantitive trait loci (sQTL) using Yoruba LCL samples as a motivating example. I build upon this work in Chapter 3. The publication provides a comprehensive resource for human structural variants (SVs).

CHAPTER 2

BENCHMARKING SEQUENCING METHODS AND TOOLS THAT FACILITATE THE STUDY OF ALTERNATIVE POLYADENYLATION

2.1 Abstract

Alternative cleavage and polyadenylation (APA), an RNA processing event, occurs in over 70% of human protein-coding genes. APA results in mRNA transcripts with distinct 3' ends, particularly found in 3' UTRs, which harbor regulatory elements that can impact mRNA stability, translation, and localization. APA can be profiled using a number of established computational tools that infer polyadenylation sites from standard RNA-seq datasets. Here, we benchmarked such cutting-edge short-read tools -- TAPAS, QAPA, DaPars2, GETUTR, and APATrap -- that take standard, short-read RNA-seq as input in their ability to identify polyadenylation sites and quantify polyadenylation site usage against 3'-Seq, a specialized RNA-seq protocol that enriches for reads at the 3' ends of genes, and Iso-Seq, a PacBio single-molecule full-length RNA-seq method. We demonstrate that 3'-Seq and Iso-Seq are able to identify and quantify the usage of polyadenylation sites more reliably than computational tools that use short-read RNA-seq as input. However, we find that running one such tool, QAPA, with a set of polyadenylation site annotations derived from small quantities of 3'-Seq or Iso-Seq can reliably quantify variation in APA across samples and genotypes, as demonstrated by the successful mapping of alternative polyadenylation quantitative trait loci (apaQTL). We envision that our analyses will shed light on the advantages of studying APA with more specialized sequencing protocols, such as 3'-Seq or Iso-Seq, and the limitations of studying APA with short-read RNA-seq. We provide a computational pipeline to aid in the identification of APA events using Iso-Seq data.

2.2 Introduction

Although the human genome only harbors about 20,000 protein-coding genes, the human transcriptome encodes ten times that number, or 200,000, of unique transcripts [35]. Over the last two decades, it has become increasingly apparent that RNA processing events, such as alternative splicing, alternative transcription start site usage, and alternative polyadenylation, are drivers of the human transcriptome's astonishing complexity, allowing single genes to encode a repertoire of transcript isoforms [45]. Alternative polyadenylation (APA), the process by which a single gene is able to produce multiple mRNA isoforms with distinct 3' ends, is a critical RNA processing event that affects the stability, localization, transport, and translation of mRNA [168, 183, 57, 79, 121, 44].

The cleavage and polyadenylation reaction is controlled by sequence elements upstream and downstream of the cleavage and polyadenylation (PAS) site. Coordinated recognition of the signal site, a hexameric A[A/U]UAAA sequence or variant thereof, $\sim$ 20-30 nucleotides upstream of the PAS site [165, 60], and a GU-rich downstream sequence element, $\sim$ 10-30 nucleotides downstream of the PAS site [44, 76, 26], is mediated by the cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) complexes. In addition to harboring alternative splice sites, genes can also harbor alternative PAS sites. By using different PAS sites, mRNA transcripts are produced with varying 3'UTR lengths, which can contain distinct cis-regulatory elements, such as miRNA binding sites or RNA binding protein sites [116, 162], and can therefore be important in the regulation of normal differentiation and development [8**?**] or mis-regulation in the context of disease [95].

The rise of next generation, high-throughput RNA sequencing (RNA-seq) [92] has allowed researchers to, without *a priori* knowledge of gene annotations, quantitatively measure gene expression, discover novel transcripts, and measure the relative abundance of distinct transcript isoforms. Studies using standard RNA-seq estimate that $\sim$ 70% of genes in the human genome harbors multiple PAS sites, most of which are localized within 3' UTRs [42]. To

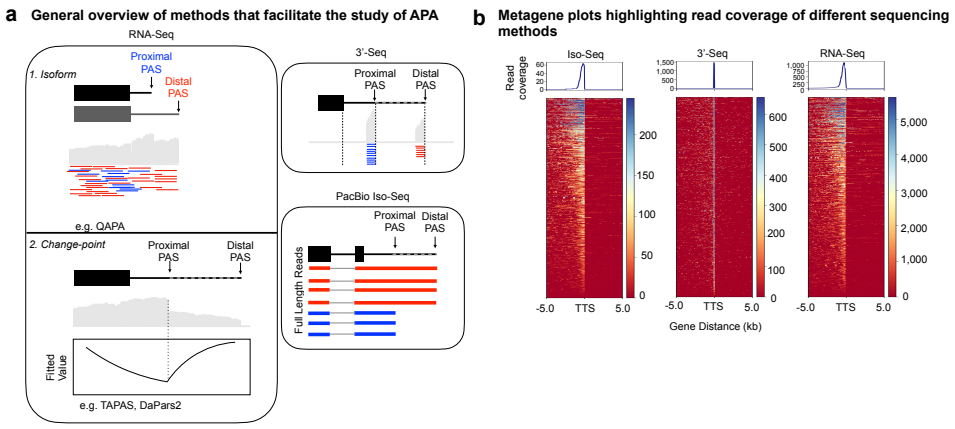


Figure 2.1: **Overview of three sequencing methods utilized to study APA.** (a) Schematic of sequencing methods - RNA-Seq, 3'-Seq, and Iso-Seq - that facilitate the study of APA. Examples of RNA-Seq methods to study APA. (b) Metagene plots showing read coverage centered around the transcription termination site (TTS) for five RNA-Seq libraries, five 3'-Seq libraries, and eight Iso-Seq libraries collected from YRI LCLs

study the effect of APA on gene regulation, a number of research groups have developed computational tools that leverage standard RNA-seq data to identify PAS sites and quantify polyadenylation site usages (PAUs). The growing number of such tools is a result of the extensive availability of short-read RNA-seq data across multiple cell types, individuals, and organisms [61, 185, 11].

Some existing approaches for studying alternative polyadenylation from short-read RNA-seq rely on estimating PAU based on transcript-level abundance [62, 84] (Fig. 2.1a). For example, QAPA calculates the relative proportion of every isoform in a gene using a combination of existing tools, namely Sailfish [132] and Salmon [131], and PAS site annotation files [62]. The use of annotations may be a drawback as PAS sites found in certain databases are often incomplete. In particular, PAS site databases may currently be missing annotation information in particular cell types or organisms of interest, biasing analyses comparing PAU across conditions. Other methods perform *de novo* identification of PAS sites by using a change-point model, which is based on a generalized likelihood ratio statistic of identifying transcript length changes [185, 11, 100] (Fig. 2.1a). For example, DaPars2 infers the location

of a single proximal PAS site within 3 UTRs [185, 100, 53]. GETUTR identifies multiple PAS sites within 3' UTRs using kernel density estimation [89]. APATrap identifies multiple PAS sites within novel 3'UTRs and 3'UTR extensions using a mean squared error model [190]. Finally, TAPAS infers PAS sites within and upstream of 3' UTRs [12].

While these methods have provided valuable insights into the landscape of APA in a myriad of biological contexts, there are a number of challenges associated with studying APA with short-read RNA-seq. Generally, estimation of isoform abundance from short-read RNA-seq is statistically challenging because short-read protocols tend to sample small portions of transcripts, and alternative transcripts often have substantial overlap [103]. Specific biases exist due to the fact that standard RNA-seq protocols include multiple PCR amplification steps during library preparation [5]. There is also bias in sequencing repetitive regions [172] and the issue of short reads not aligning uniquely within a reference genome of interest [33]. Most importantly, coverage of RNA-seq at the 3' end of mRNA transcripts is often limited, which makes estimation of the PAU particularly difficult. 3' end sequencing (3'-Seq), which enrich for reads covering the 3' ends of genes [105], overcomes the issue of limited coverage but suffers from some of the other biases associated with standard RNA-seq, such as mapping errors associated with reads derived from repetitive regions in the genome (Fig. 2.1b).

Because of the biases and analytical challenges associated with short-read sequencing protocols and their variants, we took advantage of the Pacific Biosciences (PacBio) single-molecule isoform-sequencing (Iso-Seq) [144] to more precisely identify PAS sites and quantify PAUs. We reasoned that because Iso-Seq enables sequencing through polyA tails [144] , some but not all of the biases associated with studying APA using 3'-Seq and short-read RNA-seq data would be minimized. Supporting this view, a recent study surveyed the sorghum transcriptome using single-molecule long reads, allowing for enhanced sorghum gene isoform annotation without the need for transcript reconstruction [4]. In this study, we benchmarked the ability to study APA on a genome-wide scale in humans using short-read RNA-seq-based

computational tools – TAPAS [12], DaPars2 [185, 100, 53], QAPA [62], GETUTR [89], and APATrap [190] against 3'-Seq and PacBio Iso-Seq. While there are many computational tools available that allow one to study APA, we chose these tools specifically because they leverage distinct approaches for studying APA. Notably, some of these tools leveraging annotation databases and estimating PAU based on transcript-level abundance while others make use of change-point-based detection methods. We aimed to highlight the relative advantages and disadvantages of each tool to inform the scientific community about which method may best serve study goals.

2.3 Results

2.3.1 Identification and quantification of polyadenylation sites using PacBio Iso-Seq long-read sequencing

To define the set of PAS sites to benchmark against, we compiled eight polyA-selected PacBio Iso-Seq lymphoblastoid cell line (LCL) libraries. Specifically, we generated five libraries for Yoruba (YRI) LCLs GM18501, GM18504, GM19144, GM19239, and GM19153 [58] and obtained three previously published Central European (CEU) LCL libraries for GM12878, GM12891, and GM12892 [58, 169]. High quality consensus circular sequences (CCS) were mapped to the hg19 human reference genome using minimap2 (v2.2.15) [97] separately for each of the eight libraries (Fig. 2.2a). In order to maximize power to subsequently identify PAS sites, aligned reads from the eight libraries were pooled together, resulting in a total of 2.83 million reads that were used in all downstream analyses (the "Methods" section).

Aligned reads containing polyA tails were extracted from the alignment files after performing a series of filtering steps, including filtering based on the length and adenosine composition of polyA tails and filtering for internal priming or mispriming to minimize false positive PAS identification (Fig. 2.2a). In brief, because Iso-seq reads should contain polyA

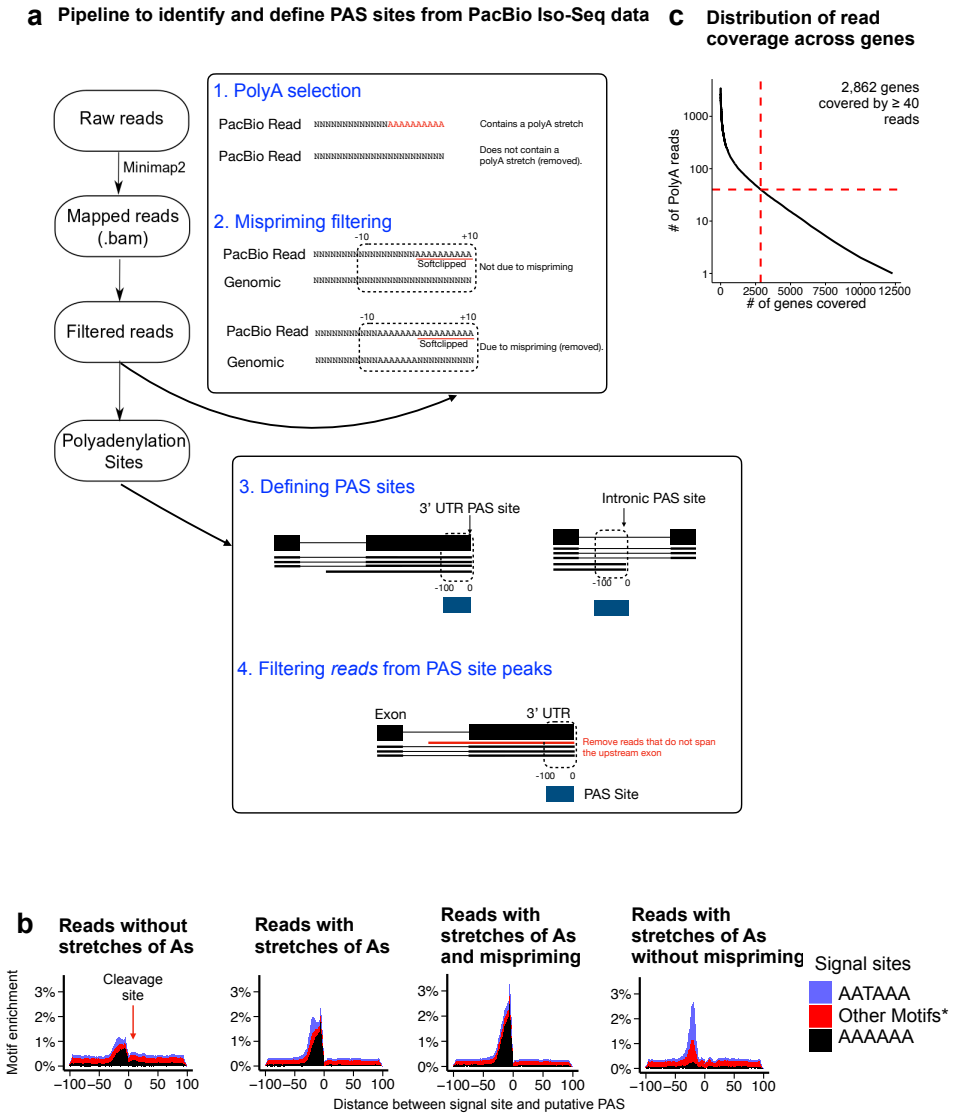


Figure 2.2: **Profiling APA with Iso-Seq.** (A) Workflow for identifying PAS sites using Iso-Seq. First, circular consensus reads were mapped to a reference genome. Next, Iso-Seq reads went through a series of two refinement steps to select for reads that contain stretches of As that are not due mispriming, resulting in a final set of reads that contain polyA tails. (B) Meta-gene barplots showing enrichment of signal site motifs $\sim$ 20-30 nucleotides upstream of the putative cleavage site within these filtered reads. There are six consensus signal site motifs, with AATAAA being the most common. Other signal sites represented as "Other Motifs" include ATTAAA, AGTAAA, TATAAA, CATAAA, and GATAAA. AAAAAA served as a negative control. (C) We were able to accurately identify PAS sites and quantify PAUs for 2862 genes, which have an Iso-Seq coverage of $\geq$ 40 reads.

tails that are not encoded in the genomic DNA, reads that contained a stretch of at least six As were retained. Moreover, reads with putative polyA tails due to mispriming were assessed by scanning the 10 nucleotides flanking either side of putative cleavage site in the genome for a stretch of at least six As and subsequently filtered out (the "Methods" section).

After filtering, we were left with 1.58 million reads with polyA tails and likely not misprimed. Using these 1.58 million reads, PAS sites were individually defined as a window between the putative cleavage site and 100 nucleotides upstream. In addition, this set of PAS sites was refined further by filtering out sites localized in the 3' UTR that did not have reads spanning an upstream exon. To define PAU, we computed the ratio of the number of reads mapping to a PAS site divided by the number of reads mapping to all PAS sites in the same gene. This resulted in a set of 27,233 PAS sites within 12,280 genes, 22,311 of which had PAUs $> 5\%$ (the "Methods" section). Our Iso-seq data analysis pipeline is available online (the "Availability of data and materials" section). We validated that the set of PAS sites that we obtained was consistent with previously defined PAS site signatures. For example, we observed enrichment of hexameric signal site motifs, such as AATAAA and ATTAAA, 20-30 nucleotides upstream of the cleavage site [165, 60, 13, 167] (Fig. 2.2b, one-sided Fisher's exact test, OR = 2.16, $p < 2.2e^6$, Fig. 2.7b) and enrichment of GT-rich sequences 10-30 nucleotides downstream of the cleavage site [76] (Fig. 2.7c, one-sided Fisher's exact test, OR = 6.03, $p < 2.2e^6$). Finally, we considered the distribution of the filtered set of reads across all genes in the genome and restricted following analyses to 2,862 genes with a read coverage of $\geq$ 40Iso-Seq reads with a polyA tail to obtain a final set of 4,446 high confidence PAS sites with PAUs $> 5\%$ (Fig. 2.2c, Fig. 2.7a).

To validate the biological utility of the aforementioned pipeline to identify PAS sites from Iso-Seq data, we applied said pipeline to previously published brain and liver Iso-Seq datasets [9] given that most other studies of APA have focused on identifying PAS sites and calculating PAUs to study differential expression of 3' UTRs across conditions, such as

across tissues. Given that read coverage across most genes in brain and liver datasets was poor (Fig. 2.8a), we restricted our analysis to 138 genes supported by at least one read that was informative with regard to the location of a PAS in the 3' UTR in both the brain and liver datasets. We observed that 30% of genes exhibited preferential usage of more distal PAS sites in the brain (at least 500 bp of distance between sites, or 20% for sites at least 1 kb apart) as compared to the liver, for which just 12% of genes used a more distal site (at least 500 bp difference, or 8% for sites at least 1 kb apart, Fig. 2.8b). This result is consistent with observations made previously that highlight a global lengthening of 3' UTRs in the brain [116].

2.3.2 Assessing PAS site features across APA detection methods

We assessed the ability of QAPA [62], DaPars2 [108, 100], TAPAS [11], GETUTR [89], and APATrap [190], 3'-Seq [106] and Iso-Seq to recapitulate known features of PAS sites. In brief, the five computational tools that leverage RNA-seq work as follows: QAPA extracts 3' UTRs for all genes from GENCODE. In addition, QAPA incorporates 3' UTR and PAS site annotation information from GENCODE [65] and the PolyASite database [71], respectively. Alternatively, a user may provide custom PAS site annotations. QAPA will then quantify PAUs by applying Sailfish [132] to resolve RNA-seq reads that map to loci containing multiple transcript isoforms. DaPars2 is a method that identifies PAS sites de novo and quantifies PAU without annotations. DaPars2 first identifies a distal PAS site for every gene based on where the RNA-seq coverage ends. From this, DaPars2 assumes that a single proximal PAS site exists, and it detects this proximal PAS site as an optical fitting point that can best explain a localized dip in read-density. DaPars2 then quantifies the PAU of proximal and distal PAS sites by adding read counts. TAPAS extracts all 3' UTRs in a gene according to a genome annotation. It then estimates the read coverage of every 3' UTR, which is given as input to the time-series data Pruned Exact Linear Time (PELT) algorithm [88] to infer

change points in a gene where the read coverage increases or decreases the most. TAPAS then filters all change points to define a true set of PAS sites, and PAUs are quantified as previously described [171]. GETUTR extracts reads that map to annotated 3' UTRs from a reference genome, makes a density function of RNA-seq data using kernel density estimation with a Gaussian kernel, and identifies PAS sites after using techniques that smooth read coverage. APATrap uses a mean squared error model to identify PAS sites. Finally, as mentioned previously, 3'-Seq is a method that enriches for reads covering the 3' end of genes.

In total, we ran TAPAS, DaPars2, QAPA, GETUTR, and APATrap using 89 RNA-seq samples from YRI LCLs as input and used our in-house pipelines (the "Methods" section) to process 54 3'-Seq YRI LCL samples and the aforementioned eight Iso-Seq YRI LCL samples. We compared the overall number of PAS sites identified by these five methods and identified 26,545, 22,062, 14,251, 46,169, 12,555, 32,286, and 22,311 PAS sites defined by TAPAS, DaPars2, QAPA, GETUTR, APATrap, 3'-Seq, and Iso-Seq, respectively (PAU $>$ 5 %). In addition, we observed that many PAS sites, regardless of which method they were defined by, were previously annotated in the database PolyA_DB 3 [176] (Fig. 2.3a).

Greater than half of the 12,280 genes expressed in LCLs harbor multiple PAS sites, or undergo APA, with 79.8% harboring $\geq$ 2 PAS sites, as defined by Iso-Seq (Fig.2.3a, 9,799 genes). Of note, although DaPars2 defines at most two PAS sites per gene, we found that 8% of genes harbored more than two PAS sites as defined by DaPars2 (Fig. 2.3b). This slight discrepancy is due to the re-assignment of PAS sites to genes using hg19 RefSeq gene annotations in order to be able to consistently compare PAS sites across TAPAS, DaPars2, QAPA, GETUTR, APATrap, 3'-Seq, and Iso-Seq in subsequent analyses.

Next, we wanted to assess the distribution of PAS sites across genic locations. We observed that while all methods agreed that most PAS sites are localized within 3' UTRs of genes, 3'-Seq identified a substantial fraction of PAS sites in introns as well (Fig. 2.3c,

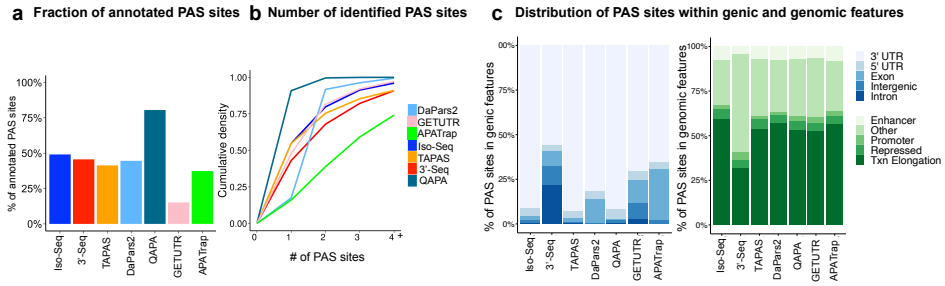


Figure 2.3: **Features of PAS sites.** (a) Barplot showing the percentage of PAS sites annotated in the PolyA_DB 3 database [176]. (b) Cumulative density of the number of identified PAS sites identified. (c) Barplot representing the genic location - 3' and 5' UTRs, introns, exons, and intergenic regions - of PAS sites as defined by HOMER [69] (left), and the genomic locations of PAS sites defined using ChromHMM annotations [50] (right). For the latter, the four annotations represented are enhancer, promoter, repressed, and transcription (txn) elongation. Eleven other annotations were collapsed together as "Other" (see the "Methods" section).

22%). This is consistent with the notion that 3'-Seq is more sensitive to APA events with low PAUs. Indeed, PAS sites in introns were used significantly less frequently than PAS site in 3' UTRs (68% and 24% of PAS sites in introns and 3'UTRs, respectively, with PAUs 20%, two proportions Z-test, $\chi^2 = 4329.5$, $p = 2.2e^{16}$). We also observed conservation upstream of the PAS site more frequently for those in 3' UTRs than those in introns or other genic locations, which is consistent with previous findings (Fig. 2.9a) [10]. Notably, most PAS sites (range of 31.4-59.3%) defined by all methods were associated with transcription elongation (59.3% of Iso-Seq PAS sites were associated with transcription elongation and 25.1% were associated with other chromatin features, two proportions Z-test, $\chi^2 = 6.221$, $p = 6.3e^3$), highlighting the importance of local chromatin architecture in PAS site selection (Fig.2.3d), as documented previously [145]. Altogether, these results suggest that all methods define PAS sites with at least one established PAS site signature, with Iso-Seq and 3'-Seq identifying more PAS sites with multiple PAS site signatures than the standard RNA-seq methods.

2.3.3 Benchmarking short-read methods against PacBio Iso-seq to identify PAS sites and quantify PAUs

To fairly and directly compare PAS sites defined by TAPAS, DaPars2, QAPA, and 3'-Seq with PAS sites defined by Iso-Seq, we restricted our analyses to the 2,862 genes with a read coverage of ≥ 40 Iso-Seq reads. Of the 4,446 Iso-Seq PAS sites (PAU $> 5\%$), 78.7% were also defined as 3'-Seq PAS sites (Fig. 2.4a, 3,500 PAS sites recovered). In comparison, TAPAS, DaPars2, QAPA, GETUTR, and APATrap were able to recover fewer Iso-Seq PAS sites, at most, 56.6% (Fig. 2.4a, TAPAS).

To assess if there were observable differences in PAUs between PAS sites within the same gene, we restricted to two PASs within the 3' UTR of every gene within our set of 2862 genes (≥ 40 Iso-Seq reads), the furthest upstream (i.e., proximal) and furthest downstream PAS site (i.e., distal). Interestingly, we observed little difference in proximal and distal PAUs identified by Iso-Seq and 3'-Seq (Fig. 2.9c). In contrast, TAPAS, DaPars2, QAPA, GETUTR, and APATrap exhibited significant PAU differences between proximal and distal PAS sites (Fig. 2.9c, prop.test, P = $1.088543e^{44}$). This was expected given that different sequence isoforms can contain a significant amount of sequence overlap, and short RNA-seq reads assigned may be assigned to the incorrect isoform. Therefore, short-read-based methods may overestimate proximal PAU. Indeed, TAPAS and DaPars2 identified a significant number of distal PAS sites with lower PAUs; 74.7% of distal PAS sites identified by TAPAS exhibited PAUs $\leq$ 50%, and 61.6% of distal PAS sites identified by DaPars2 exhibited PAUs $\leq 50\%$ (Fig. 2.9c). Interestingly, QAPA identified slightly more proximal PAS sites with lower PAUs (Fig. 2.9c). This may be a result of the fact that QAPA was run with additional annotation information as compared to TAPAS and DaPars2.

Next, we compared PAUs across the different methods at the gene-level. More precisely, if two methods called the same PAS sites for a specific gene, we computed the difference in their usages and summed across the differences for all PAS sites within that gene (Fig. 2.4b,

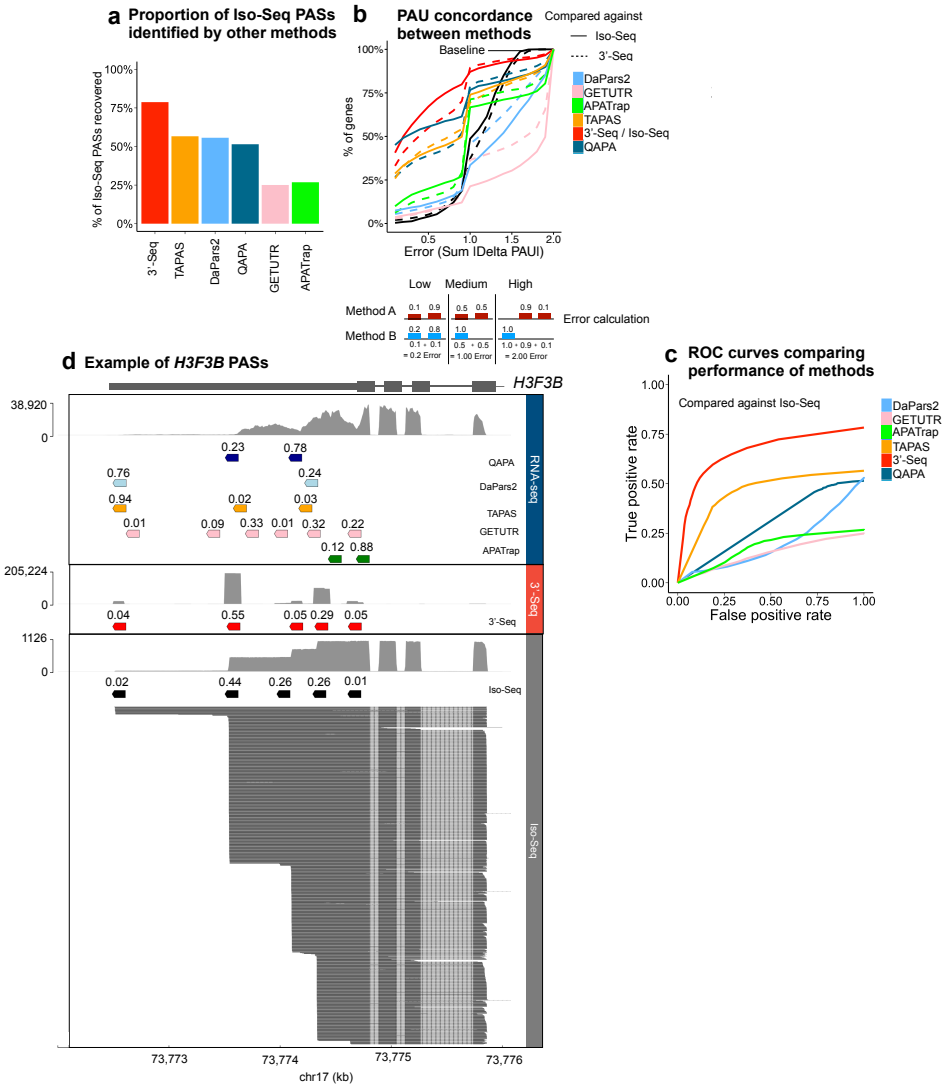


Figure 2.4: **Comparing PAS site identification and PAU quantification across methods.** (A) Barplot showing the proportion of Iso-seq PAS sites across 2,862 genes that were identified by short-read sequencing methods, 3'-Seq, QAPA, DaPars2, and TAPAS. These methods are able to identify, at best, $\sim$ 75% PAS sites identified by Iso-Seq. (B) Comparison of PAU calls across methods. Error(Sum |Delta PAU|) refers to the concordance in calls between two methods, A and B (as outlined in the schematic). The solid and dotted lines represent comparison of all methods against Iso-Seq, and 3'-Seq, respectively. (C) Receiver operating characteristic (ROC) curves. True positives are instances in which Iso-Seq with PAUs $> 5\%$ have analogous PAS sites defined by other methods with PAUs $> 0\%$. False positives are PAS sites defined by other methods with PAUs $> 5\%$, but lack analogous PAS sites defined by Iso-Seq. (D) RNA-seq, 3'-Seq, and Iso-seq coverage track at the *H3F3B* 3'UTR, showing PAS sites and PAUs ($> 1\%$) identified by Iso-Seq and the short-read methods.

Fig. S4b). We defined this as the amount of "error" or difference between PAUs estimated by different methods (Fig. 2.4b). Examples of low error include cases in which two methods define the same PAS sites per gene but might estimate PAUs to be slightly different. In contrast, examples of high error include cases in which two methods may define completely different PAS sites per gene. When comparing 3'-Seq, TAPAS, DaPars2, QAPA, GETUTR, and APATrap against Iso-Seq (Fig. 2.4b, Fig. 2.10b "Concordance with Iso-Seq"), 3'-Seq was most concordant. GETUTR was least concordant, with 78.6% of genes tested having an error > 1.0, suggesting a large discrepancy between GETUTR-defined PAUs and Iso-Seq-defined PAUs. All standard RNA-seq-based tools, TAPAS, DaPars2, QAPA, GETUTR, and APATrap, were similarly concordant with 3'-Seq (Fig. 2.4b, Fig. 2.10b "Concordance with 3'-Seq").

We generated a receiver operating characteristic (ROC) curve to highlight the tradeoff between sensitivity and specificity of 3'-Seq, TAPAS, DaPars2, QAPA, GETUTR, and APATrap as compared to the Iso-Seq sites, which, for the purpose of this analysis, were the ground truth (Fig. 2.4c). We did not simulate synthetic datasets for this analysis as, to date, there are very few methods that can simulate Iso-Seq data, and the methods that do exist simulate reads lacking polyA tails, rendering them uninformative for the study of APA. We defined a true positive as an instance in which a method defines an analogous PAS site that overlaps the Iso-Seq PAS by at least a single base with a PAU $> 5\%$. In contrast, we defined a false positive to be an instance in which a PAS site is defined by 3'-Seq, TAPAS, QAPA, GETUTR, APATrap, or DaPars2 with a PAU $> 5\%$, but there does not exist an overlapping PAS site defined by Iso-Seq. Overall, 3'-Seq outperformed all other methods, as measured by the area under the curve (Fig.2.4c, AUC = 0.66). In contrast, the AUCs for TAPAS, QAPA, DaPars2, GETUTR, and APATrap were 0.46, 0.50, 0.20, 0.17, and 0.20, respectively. We repeated this analysis with the 3'-Seq PAS sites as the ground truth and observed similar results (Fig. 2.10e).

To showcase the complexity of PAS site identification and PAU quantification, we highlight *H3F3B* as an example in which Iso-Seq and 3'-Seq identified five PAS sites with (PAUs $> 1\%$), with comparable PAUs (Fig. 2.4d). In contrast, the standard RNA-seq based methods identified overall fewer PAS sites, on average, with PAUs that did not agree with Iso-Seq and 3'-Seq.

2.3.4 Evaluating the study of inter-individual variation in PAS site choice using different sequencing methods

We have demonstrated thus far that 3'-Seq PAS site identification and PAU quantification is superior to that of RNA-seq-based methods. Nevertheless, a large number of RNA-seq datasets are publicly available, which can be readily used to study APA. Therefore, we sought to evaluate the possibility of combining short-read RNA-seq with a set of PAS site annotations derived from small quantities of 3'-Seq or Iso-Seq to study variation in APA across samples.

As a possible test-case, we set out to use human population-scale RNA-seq data alongside 3'-Seq and Iso-Seq PAS site annotations to study the impact of genetic variation on PAU. While previous studies have focused on comparing PAU across conditions, such as cell types, tissues, or species, we note that studying the impact of genetic variation on PAU is simply another type of comparison of PAU across conditions. In this case, there are three conditions, each one a possible genotype. Moreover, to date, very few studies have quantified APA in a human population samples and detected genetic variants implicated in genome-wide APA variation [100, 122], although this may change in the future.

Given that studying the impact of genetic variation on variation in PAU across diverse human populations necessitates the use of well-powered datasets with samples collected from many individuals, and the fact that many such datasets are based on standard, short-read RNA-seq protocols, we first wished to compare the ability to call APA quantitative

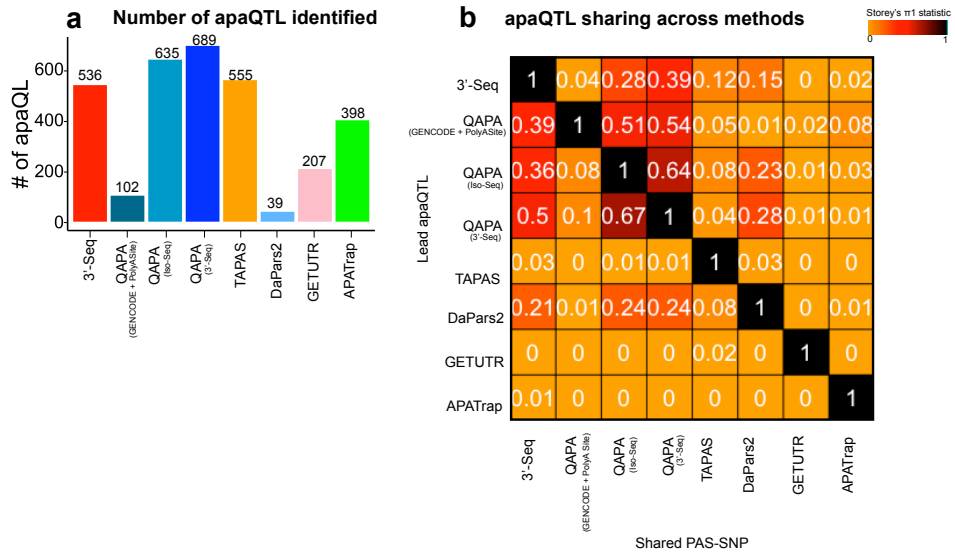


Figure 2.5: **Studying the impact of genetic variation on variation in PAS site choice using different sequencing methods.** (A) Barplot showing the number of apaQTL (FDR $< 10\%$) identified across the short-read RNA-Seq-based methods and 3'-Seq. (B) Heatmap showing sharing of apaQTLs across sequencing methods using Storey's π_1 statistic.

trait loci (apaQTL), which link variations in PAU to genotype, accurately using PAS sites defined by TAPAS, DaPars2, QAPA, GETUTR, and APATrap by benchmarking against apaQTL we identified based on PAS sites defined by 3'-Seq. We next ascertained whether the reliability and reproducibility of apaQTL called using RNA-seq-based tool PAS sites as input, in particular QAPA, could be improved when given custom PAS site annotations based on small quantities of specialized datasets, such as 3'-Seq and Iso-Seq.

In order to map apaQTL, we restricted our analysis to samples with genotype information: 87 YRI LCLs for which we had access to short-read RNA-seq data and 51 YRI LCLs for which we had access to 3'-Seq data (the "Methods" section). We did not map apaQTL using Iso-Seq data because data from only 8 individuals were available to us, much fewer than the 50 individuals that are typically required for QTL analysis.Briefly, per individual, we defined PAUs for each PAS site defined by every APA method separately, as described previously, now including PAS sites with PAUs $< 5\%$. We quantile normalized these ratios

and tested for the association between PAU and single nucleotide polymorphisms (SNPs) within 25 kb of the associated PAS site using FastQTL [128]. Significant SNP-PAS pairs were defined as apaQTL (FDR <10%). We were able to identify hundreds of apaQTL, but the exact number of apaQTL mapped varied greatly between methods and tools (Fig.2.5a). In particular, we noted a substantial increase in the number of apaQTL called using PAS sites defined by QAPA run with Iso-Seq or 3'-Seq annotations as compared to QAPA run with the combination of GENCODE and PolyASite annotations, which is the default (Fig. 2.5a, 635 and 685 versus 102 apaQTL called by QAPA run with Iso-Seq, 3'-Seq, and GENCODE and PolyASite annotations, respectively). The number of apaQTL called using PAS sites defined by QAPA run with Iso-Seq or 3'-Seq annotations is comparable in number to that of 3'-Seq (Fig. 2.5a, 635 and 685 versus 536 apaQTL called by QAPA run with Iso-Seq and 3'-Seq annotations, respectively). We observed enrichment of apaQTL near cleavage sites, suggesting that all methods and tools were able to identify apaQTL that likely enhance or disrupt recognition of signal sites (Fig. S5b).

To assess what fraction of apaQTL called by the various tools could be recapitulated by other tools, we estimated sharing of apaQTLs using Storey's π_1 statistic and restricted to PAS-SNP pairs within 3' UTRs, as some of the short-read RNA-seq-based tools do not identify PAS sites upstream of 3' UTRs. Interestingly, very few apaQTL called using PAS site defined by TAPAS, GETUTR, and APATrap were shared with apaQTL called by other methods (Fig. 2.5b), suggesting potential false positives. We also observed that while only 4% of apaQTL identified by 3'-Seq were shared with those identified by QAPA run with GENCODE and the PolyASite database annotations, 28% and 39% of apaQTL were shared by QAPA run with Iso-Seq or 3'-Seq PAS site annotations, respectively (Fig. 2.5b). For example, rs7029002 (C>G) is an apaQTL identified using 3'-Seq data that was shared with QAPA run with Iso-Seq and 3'-Seq PAS site annotations. As apaQTL represent associations between genotype and PAU, at this locus, individuals with more G alleles at rs7029002

exhibit higher PAUs associated with the PAS site at the end of the 3' UTR of the *DDX58* gene as compared to individuals with more C alleles (Fig. 2.6). While this apaQTL is also significant when called using PAU quantifications from TAPAS and DaPars2, significance is greatly diminished. Moreover, the estimated effect size is reversed for DaPars2, meaning that in this case, individuals with C alleles, instead of G alleles, exhibit higher PAU of the PAS site at the end of the 3' UTR (Fig. 2.6). Interestingly, rs7029002 is upstream of the PAS site it is associated with, suggesting that it is tagging a genetic variant that likely disrupts recognition of the cleavage site, either directly or indirectly. Some apaQTL were exclusively identified by 3'-Seq, such as rs72836634 near the gene *CASC3* (Fig. 2.12). Nevertheless, a large fraction of apaQTL called using 3'-Seq could be identified using QAPA with Iso-Seq or 3'-Seq annotations (Fig. 2.5b, $\pi_1 = 0.28$ and 0.39), suggesting that running QAPA with Iso-Seq or 3'-Seq annotations derived from a small number of individuals is a reasonable alternative to performing 3'-Seq in all individuals for apaQTL mapping.

Lastly, we quantified sharing of apaQTL called using PAS-defined by the different tools with expression quantitative trait loci (eQTL), which can serve as a proxy for function. For example, shared apaQTL and eQTL include cases in which one PAS may provide more stabilility to a transcript over another, preventing the transcript from being subject to degradation. In such an example, high gene expression might serve as a proxy for transcript stability. Overall, we observed that sharing of apaQTL with associated gene-SNP pairs was indeed relatively high (Fig. 2.11d, π_1 between 0.09 and 0.36). In particular, sharing between eQTL and apaQTL identified using 3'-Seq ($\pi_1 = 0.36$) is very similar to that between eQTLs and apaQTLs identified using QAPA run with 3'-Seq ($\pi_1 = 0.39$) or run with Iso-Seq annotations ($\pi_1 = 0.36$) (Fig. 2.11d). This observation suggests that the apaQTL identified using these three methods are likely to have similar functional impacts on gene regulation.

Overall, these observations suggest that, indeed, under circumstances in which APA-related specialized datasets cannot be generated for a large sample of individuals, QAPA, run

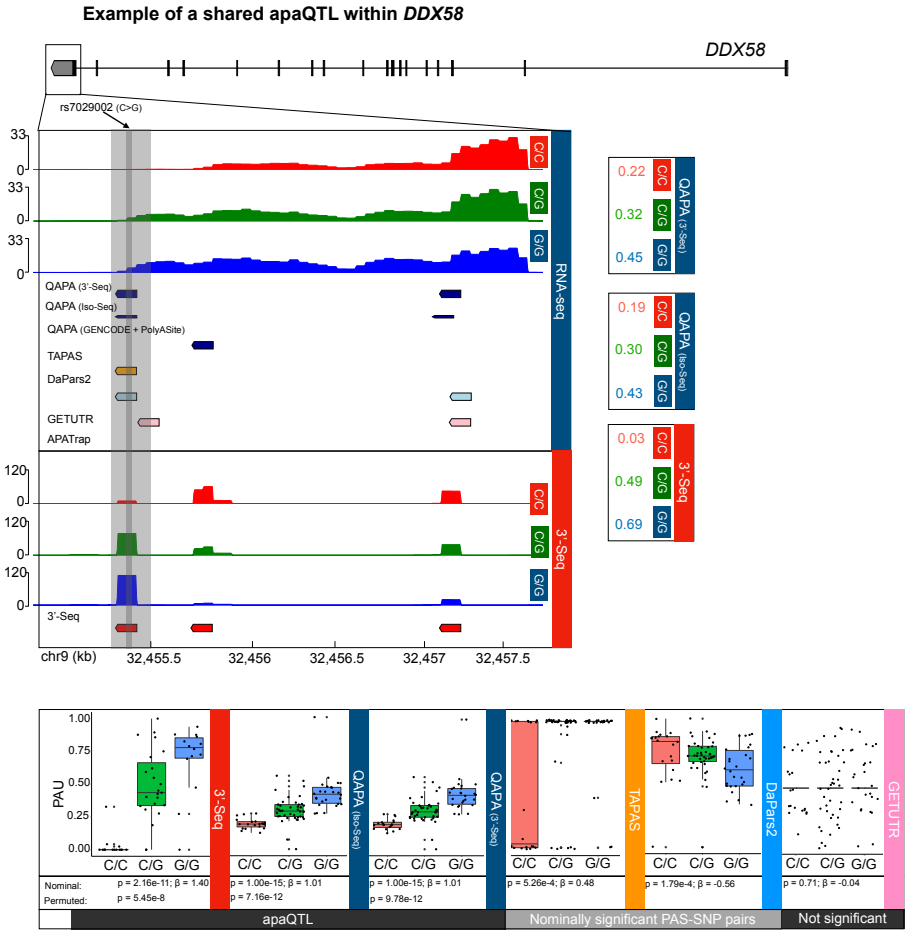


Figure 2.6: **Example of a shared apaQTL in the gene DDX58 defined by 3'-Seq and shared by QAPA run with Iso-Seq or 3'-Seq PAS site annotations.** We highlight a gene track displaying read coverage and PAS sites. The gray vertical track directly below rs7029002 represents the position of the strongest apaQTL SNP, and the surrounding gray track represents the most strongly affected PAS site. PAUs for every PAS site were stratified by genotype, as shown on the right, in which individuals with the G allele have increased PAU of the highlighted, gray PAS site. The bottom highlights boxplots of the PAU at this PAS site, stratified by genotype and APA detection method.

with custom PAS site annotations derived from small quantities of such specialized datasets, recapitulates PAS sites and apaQTL that would otherwise be identified using population-scale 3'-Seq data.

2.4 Discussion

Short-read RNA-seq has become central in the assessment of transcriptional and post-transcriptional gene regulatory mechanisms, such as APA, which contributes substantially to the amount of diversity in the human transcriptome and proteome by increasing the number of isoforms produced through differences in PAS site selection. However, given the limited size of RNA-seq sequence fragments and inherent complexity of the human transcriptome, it remains difficult to accurately reconstruct full-length transcripts with short-read RNA-seq. 3'-Seq, a specialized RNA-seq protocol that enriches for reads at the 3' ends of genes is a better, well-established alternative for the study of APA. Moreover, single-molecule long-read RNA-seq, such as PacBio Iso-Seq, offers a considerable advantage over short-read sequencing to more precisely identify PAS sites and quantify PAUs across mammalian transcriptomes because this protocol allows for capture of full-length transcripts, including polyA tails, thus obviating the need for transcript reconstruction entirely.

In this study, we identified 22,311 PAS sites (PAU $>$ 5%) across 12,280 genes, 38.7% of which are novel, using Iso-Seq data derived from eight LCL samples. We observed that APA detection methods, such as those that take short-read RNA-seq as input, including as TAPAS, DaPars2, QAPA, GETUTR, and APATrap, as well as 3'-Seq, were able to identify comparable numbers of PAS sites. Importantly, PAS sites identified by all methods exhibited well-characterized features of PAS sites, including enrichment of signal site motifs upstream of cleavage sites, enrichment within 3' UTRs, and association with transcription elongation.

We benchmarked the ability to study APA using all tools against 3'-Seq and Iso-Seq. We estimated that 78.6% of PAS sites identified by 3'-Seq overlap with Iso-Seq-defined PAS

sites whereas, at best, 56.6% PAS sites identified by one of the RNA-seq-based methods, QAPA, overlap with Iso-Seq-defined PAS sites. Moreover, as expected, there is reasonable concordance in identification of PAS sites and estimation of PAUs between 3'-Seq and Iso-Seq, with some differences. This is in contrast to the greater discordance between RNA-seq-based tools and Iso-Seq, likely because PAS site identification and PAU quantification among the RNA-seq-based tools can be highly variable. Overall, this suggests that researchers should carefully assess which RNA-seq-based tools might serve them best based on the exact biological questions they may be interested in answering. Moreover, 3'-Seq should be the method of choice for studying APA when such data can be generated or are available.

We acknowledge that it is not necessarily practical or cost-effective to generate specialized datasets to study APA, especially given that a plethora of short-read RNA-seq already exist for a large number of samples. Through our analysis of inter-individual variation in APA as a test-case, it is apparent that QAPA, an isoform-based RNA-seq method to study APA, paired with PAS site annotations derived using small quantities of specialized sequencing data, such as 3'-Seq and Iso-Seq, may offer a considerable advantage in studying APA in a cost-effective manner in the near term until it becomes more accessible and inexpensive to study APA extensively using full-length, long-read sequencing.

2.4.1 Conclusions

This study demonstrates that current methods to study RNA processing events, such as APA, with short-read RNA-seq data suffer from limitations. However, combining large quantities of RNA-seq data with small quantities of specialized data, in this case, 3'-Seq or Iso-Seq, strikes an attractive balance between affordability and accuracy in the study of APA.

2.5 Methods

2.5.1 Cell culture and RNA sample preparation

We cultured 5 Epstein-Barr Virus transformed lymphoblastoid cell lines (LCLs) at 37 C and 5% CO2. These LCLs -- GM18501, GM18504, GM19144, GM19239, and GM19153 -- were derived from the Yoruba (YRI) individuals from the International HapMap Consortium. The Coriell Cat #:Research Resource Identifiers (RRIDs) are GM18501:CVCL_P458, GM18504:CVCL_P460, GM19144:CVCL_P525, GM19239:CVCL_9634, and GM19153:CVCL P531.

These lines were authenticated and tested for mycoplasma contamination. Cell culture and RNA extraction were performed as described previously [122]. In brief, cells were grown in a glutamine depleted RPMI [RPMI 1640 1X from Corning (15-040 CM)] with 15% FBS, 2 mM GlutaMAX (from gibco (35050-061), 100 IU/mL Penicillin, and 100 ug/mL Streptomycin. The lines were passaged 3 times, maintained at 8105 cells, and grown to a concentration of 1106 cells per mL before RNA extraction, which was performed as described previously [122]. In brief, cells from each line were spun down and pelleted at 200 g at 500 RPM at 4C for 2 min, washed with cold phosphate-buffered saline (PBS), and spun down again before aspirating the PBS. RNA was extracted using the miRNeasy kit (Qiagen) according to the manufacturer's instructions, including the DNase step to remove potentially contaminating genomic DNA.

2.5.2 Long-read RNA-sequencing data mapping, filtering, and quality control

We processed a total of 8 polyA-selected PacBio Iso-Seq LCL libraries [54] Five SMRT bell libraries were generated for the aforementioned YRI LCLs, GM18501, GM18504, GM19144, GM19239, and GM19153, as per the PacBio Iso-Seq protocol described previously [135]. In

brief, cDNA synthesis was performed in triplicate, with each reaction starting with 800-1000 ng of total RNA. The samples were sequenced using 4 SMRTcells. We generated consensus circular sequences (CCS), removed primers, demultiplexed samples, and converted to fastqs [154]. In addition to the 5 YRI LCL samples we generated, we leveraged previously published the Central European (CEU) LCL libraries, GM12878, GM12891, and GM12892 (NCBI SRA SRP036136) [170].

Reads were mapped to the hg19 human reference genome using minimap2 (version 2.2.15) separately for every library [97], using the specific parameters *minimap2 -ax splice -uf -C5 hg19.fa ¡file>.fastq > ¡file>.sam*. In order to increase power to call PAS sites, aligned reads from the eight libraries were pooled together.

To identify Iso-Seq reads that capture cleavage and polyadenylation events, we searched for reads that contained stretches of adenosines (i.e., polyA tails). PolyA stretches needed to be located immediately after the 3' end of the alignment (i.e., starts at the base within the read that does not map to the hg19 reference genome, which is otherwise known as the portion of the read that is "softclipped"). We assessed if the softclipped portion of every read contained a stretch of adenosines. We retained the reads if their softclipped segments were < 20 nucleotides in length and were composed of 95% adenosines. Moreover, if the length of the softclipped segment of a read was ≥ 20 nucleotides, we assessed if the first 20 nucleotides of the softclipped segment was composed of 80% adenosines and if the following 20 nucleotides of the softclipped segment was composed of 95% adenosines and retained these reads as containing a stretch of adenosines.

Next, reads with stretches of adenosines were filtered for internal priming or mispriming using an approach similar to what has been described previously [102]. In brief, we extracted 20 base pairs of the genomic sequence flanking the cleavage site (i.e., 10 nucleotides upstream and 10 nucleotides downstream of the base at which the softclipping segment began) and discarded reads that contained 6 out of 10 adenosines upstream or 6 out of 10

adenosines downstream. We considered this final set of reads as having reliable polyA tails, and therefore, we used this set for downstream analyses.

To ensure the validity of our filtering steps, we verified that the set of final reads showed enrichment of hexameric polyadenylation signals (e.g., AAUAAA), as described previously [26, 13, 102]. In addition, we also verified for enrichment of other sequence elements that are known to play an important role in correct cleavage site recognition, namely a downstream element that contains GU-rich sequences [44, 76, 26].

2.5.3 Iso-Seq PAS site identification and PAU quantification

Putative PAS sites were defined as the ends of the mapped portion (i.e. the cleavage site) of reads and 100 nucleotides upstream. We then refined the set of reads used to define every putative PAS site by filtering out reads that did not map to annotated 3' UTRs and that did not also span an upstream exon. We then refined this set of PAS sites by restricting to those in annotated genes using the annotatePeaks.pl script (HOMER v4.11) [69]. This script also annotates with information about the genic location, such as the 5' or 3' UTR, intron, and exon of a peak, or in this case a PAS site. We restricted to PAS sites that fell within genes with ≥ 40 Iso-Seq reads with a polyA tail. PAUs were quantified by counting the number of reads that ended at a particular PAS site divided by the total number of reads that ended at any PAS site within the same gene. For downstream analyses, we restricted to PAU > 5 % (Additional file 2).

2.5.4 3'-Seq PAS site identification and PAU quantification

We used 3'-Seq data that were generated from 54 LCLs previously (NCBI Sequence Read Archive Accession (SRA) SRP223759, total fraction) [122]. Reads were aligned to the hg19 reference genome using STAR v2.6 [46]. Next, reads were filtered for internal priming or mispriming by locating a stretch of 6 adenosines in a 22 nucleotide window surrounding

the cleavage site (10 nucleotides upstream and 12 nucleotides downstream), similar to [105]. As was done in for the Iso-seq data, we evaluated enrichment of AAUAAA upstream of the cleavage site. From this final set of reads, peaks were identified as described previously [105]. In brief, peaks were identified by convolving the read coverage with the second derivative of a Gaussian filter such that the lowest convolved read coverage value was defined as the peak center. The peak was then extended 100 nucleotides upstream. Peaks supported by fewer than an average of 5 reads were discarded. This set of peaks was then refined by restricting to those in annotated genes as per the annotatePeaks.pl script from HOMER [69], and PAUs were quantified as described previously for Iso-Seq PAS sites. For downstream analyses, we restricted to PAU $> 5\%$ (Additional file 3).

2.5.5 Short-read RNA-sequencing and 3' end sequencing data processing and mapping

Standard, short-read RNA-seq data for 89 LCLs (NA18486, NA18487, NA18488, NA18489, NA18498, NA18499, NA18500, NA18502, NA18505, NA18508, NA18510, NA18511, NA18517, NA18519, NA18520, NA18858, NA18861, NA18867, NA18868, NA18870, NA18873, NA1897, NA18907, NA18908, NA18909, NA18910, NA18912, NA18916, NA18917, NA18923, NA18933, NA18934, NA19093, NA19095, NA19096, NA19098, NA19099, NA19102, NA19107, NA19108, NA19113, NA19114, NA19116, NA19117, NA19118, NA19119, NA19121, NA19129, NA19130, NA19131, NA19137, NA19138, NA19141, NA19143, NA19144, NA19146, NA19147, NA19149, NA19150, NA19152, NA19153, NA19159, NA19160, NA19171, NA19172, NA19175, NA19184, NA19185, NA19189, NA19190, NA19197, NA19198, NA19200, NA19201, NA19204, NA19206, NA19207, NA19209, NA19210, NA19213, NA19214, NA19222, NA19223, NA19225, NA19235, NA19236, NA19247, NA19248, NA19256, and NA19257) was obtained from the GEUVADIS project (EBI ArrayExpress, under the accession E-GEUV-1). In brief, reads were mapped to the hg19 human reference genome using STARv2.6 [46]. Aligned reads were used as input for

different tools that allow for identification of PAS sites from RNA-seq data. Because QAPA is an annotation-based method, we used QAPA's pre-compiled hg19 annotation library (`https://zenodo.org/record/1222196/files/qapa_3utrs.gencode.hg19.tar.gz`), which is derived from GENCODE [65] and the PolyASite database [71] together, as was done previously [61]. In addition, we also ran QAPA with two other annotation files, namely the BED files of Iso-Seq PAS sites and 3'-Seq PAS sites that we generated, separately. When running QAPA with these custom annotation files, we extended the 3' UTRs extracted from the hg19 GENCODE gene prediction annotation tables by 1 kb in order to avoid QAPA not identifying PAS sites that were present in our custom annotation files.

Moreover, all RNA-seq based tools output estimates of PAU separately for every individual. Therefore, we averaged PAU across all individuals for downstream comparisons. The PAS sites were re-annotated with HOMER as described previously, and PAUs for every gene were re-scaled to sum to 1.0 if any PAS sites were omitted because they could not be annotated by HOMER. For downstream analyses, we restricted to PAU¿5% (Additional files 4, 5, 6, 7, 8, 9 and 10).

2.5.6 Assessing the number of PAS sites within annotated PAS site databases

We used hg19 PAS site annotations derived from PolyA_DB 3 (release 3.2, August 2018) [176] to assess the proportion of PAS sites that were previously annotated.

2.5.7 Conservation analysis

In the analysis of sequence conservation, we used phyloP scores generated on the 46-way vertebrate alignment, restricting to placental mammals. These were downloaded from the UCSC Genome Browser [140].

2.5.8 Benchmarking short-read RNA-seq tools and 3'-Seq against Iso-Seq

To assess the concordance between PAS site location and PAU quantification defined by the RNA-seq based tools and 3'-Seq as compared to those defined by Iso-Seq, we restricted to PAS sites that fell within the set of 2862 genes with Iso-Seq read coverage ≥ 40. In addition, we restricted to Iso-Seq PAS sites with PAUs $> 5\%$.

We directly assessed the overlap of PAS sites called by different methods using BEDTools [142]. We defined an error metric, Error(Sum | Delta PAU |), which measures the concordance in PAU calls between two methods, A and B. This measure jointly assesses PAS site localization and PAU quantification concordance. In brief, for every gene, we summed over the differences in PAUs between all PAS sites defined by methods A and B.

2.5.9 Sensitivity and specificity analysis

ROC curves were generated to assess the sensitivity and specificity of the RNA-seq tools and 3'-Seq in accurately identifying PAS sites. Specifically, the Iso-Seq PAS sites with PAUs $>$ 5% were used as the ground truth. True positives are instances in which Iso-Seq PAS sites with PAUs $> 5\%$ have analogous PAS sites defined by other methods with PAUs $> 0\%$. False positives are PAS sites defined by other methods with PAUs $> 5\%$, but lack analogous PAS sites defined by Iso-Seq.

2.5.10 apaQTL mapping

We mapped apaQTL separately for all methods except Iso-Seq, for which we were lacking power to call QTL given our small sample size of eight individuals. For the RNA-seq methods, we removed two individuals, NA18500 and NA18908 due to low confidence in their annotated identity (a remaining total of 87 individuals). For the 3'-Seq, we removed these same two individuals as well as NA19092 due to lack of genotype information (a remaining total of 51 individuals). We analyzed all PAS sites defined by each of these methods, regardless of

PAU.

We standardized all PAU measurements across individuals and then quantile-normalized them to fit a standard normal distribution, as described previously degner2012dnase, van2015wasp. We used principal components analysis (PCA) to regress out confounders. We regressed out four PCs. To map apaQTL, we ran FastQTL and used all SNPs with MAF > 0.05 within 25kb of PAS sites [64]. As input, we used SNPs from GEUVADIS [94, 104]. A P-value from a standard linear regression was extracted from the FastQTL output for every SNP-PAS pair. In addition, the lead SNP-PAS association for every PAS site was obtained from the 1,000 permutations performed by FastQTL [27]. apaQTL were defined as SNPs from this set with FDR $< 10\%$.

2.5.11 eQTL mapping

We mapped eQTL in a fashion analogous to apaQTL, now with the molecular phenotype as gene expression instead of PAU. The same set of RNA-seq data from 87 individuals were used. The same set of SNPs, with MAF > 0.05 were used, now within 1MB of genes.

2.5.12 Estimation of QTL sharing

To estimate sharing between apaQTL mapped using PAS site derived from different methods, we used Storey's π_1 method (otherwise known as qvalue()), which considers the P-value of the lead SNP-PAS pair from method A in method B [38]. Similarly, we also estimated sharing between QTL for the molecular traits APA and gene expression in an analogous fashion in which we considered the P-value of the association between the lead SNP-PAS pair and gene expression level.

2.5.13 Data and code availability

The datasets supporting the conclusions of this article are available: the 89 YRI LCL RNA-seq dataset was generated by the GEUVADIS project and is available in the EBI ArrayExpress repository, E-GEUV-1 [46]. The 54 YRI LCL 3'-Seq dataset was generated by Mittleman et al. 2020 and is available in the NCBI Sequence Read Archive Accession (SRA) repository, SRP223759 (total fraction). The 3 CEU LCL PacBio Iso-Seq dataset was generated by Tilgner et al. 2014 and is available in the NCBI Sequence Read Archive Accession (SRA) repository, SRP036136. The 5 YRI LCL PacBio Iso-Seq dataset was generated in the current study and is available in the NCBI Sequence Read Archive Accession (SRA) repository, PRJNA762669 [154].

All reproducible scripts can be found through Zenodo. A more detailed Iso-Seq analysis pipeline is available on GitHub: https://github.com/ankeetashah/Benchmarking-APA under an MIT license. All other scripts are available upon request.

2.6 Acknowledgments

This work was completed in part with resources provided by the University of Chicago's Research Computing Center. We thank The University of Chicago Genomics Facility (RRID:SCR_019196), especially Pieter W. Faber and Mikala Marchuk, for their assistance with PacBio Iso-Seq cDNA synthesis, library preparation, and SMRT Iso-Seq. This work was supported by the US National Institutes of Health (R01GM130738 to Y.I.L and T32GM007197 to A.S.).

2.7 Acknowledgement of work performed

The work presented in this chapter was adapted from the journal article: Ankeeta Shah, Briana E Mittleman, Yoav Gilad, Yang I Li. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. Benchmarking computational tools

and methods that facilitate the study of alternative polyadenylation. *Genome Biology* 22 (1), 1-21, 2021

I would like to acknowledge the individuals who contributed to this work. The analyses presented in this chapter were conceived, generated, and visualized by Ankeeta Shah. Briana Mittleman extracted RNA.

2.8 Supplementary information

2.8.1 Supplemental figures

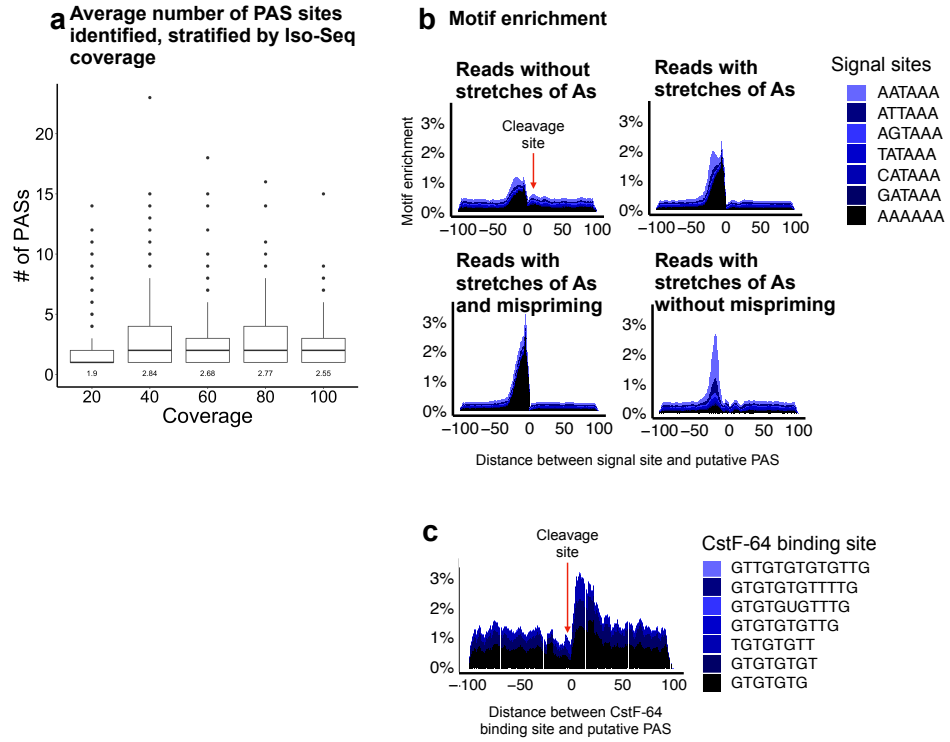


Figure 2.7: **Iso-Seq data filtering criteria for the study of APA.** (a) Boxplot showing the average number of PAS sites identified per gene, binned by coverage. We observed that genes with coverage $\geq$ 40 Iso-Seq reads consistently have, on average, 2-3 PAS sites. (b) Meta-gene plots showing enrichment of signal site motifs, AATAAA, ATTAAA, AGTAAA, TATAAA, CATAAA, and GATAAA, 20-30 nucleotides upstream of the putative cleavage site within filtered reads. AAAAAA serves as a negative control. (c) Meta-gene plot showing the enrichment of the GT-rich binding site of CstF-64 10-30 nucleotides downstream of the putative cleavage site.

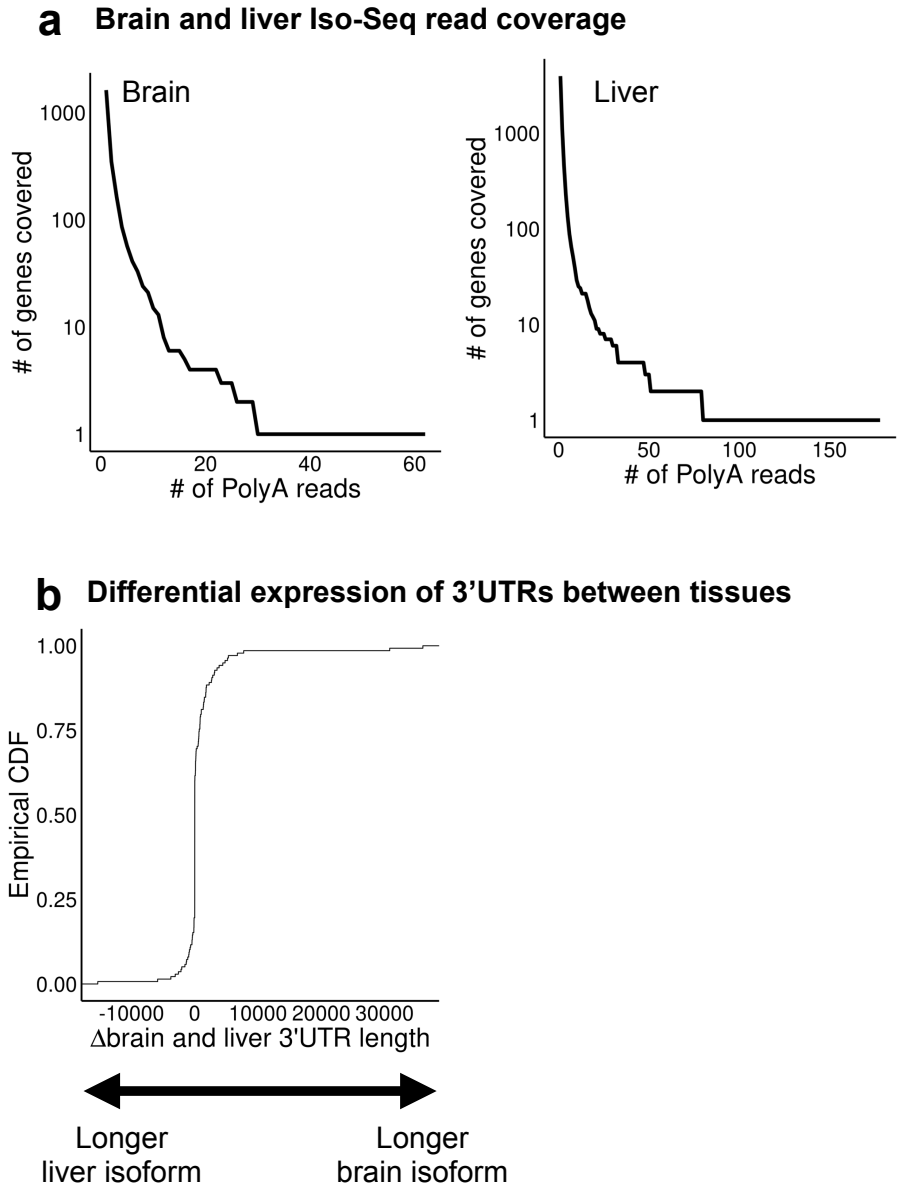


Figure 2.8: **Differential expression of alternative 3'UTRs between tissues.** (a) Read coverage supporting PAS sites in the 3'UTRs of genes derived from previously published brain and liver Iso-Seq datasets [9] (b) Among 138 genes with a PAS site supported by at least one read, we observed that 30% exhibited use of more distal PAS sites (i.e. longer 3'UTRs, 30% at least 500bp difference, or 20% for sites at least 1kb apart) as compared to liver, which exhibited increased use of more proximal sites (i.e. short 3'UTRs, at least 500kb apart, or 8% for sites at least 1kb apart).

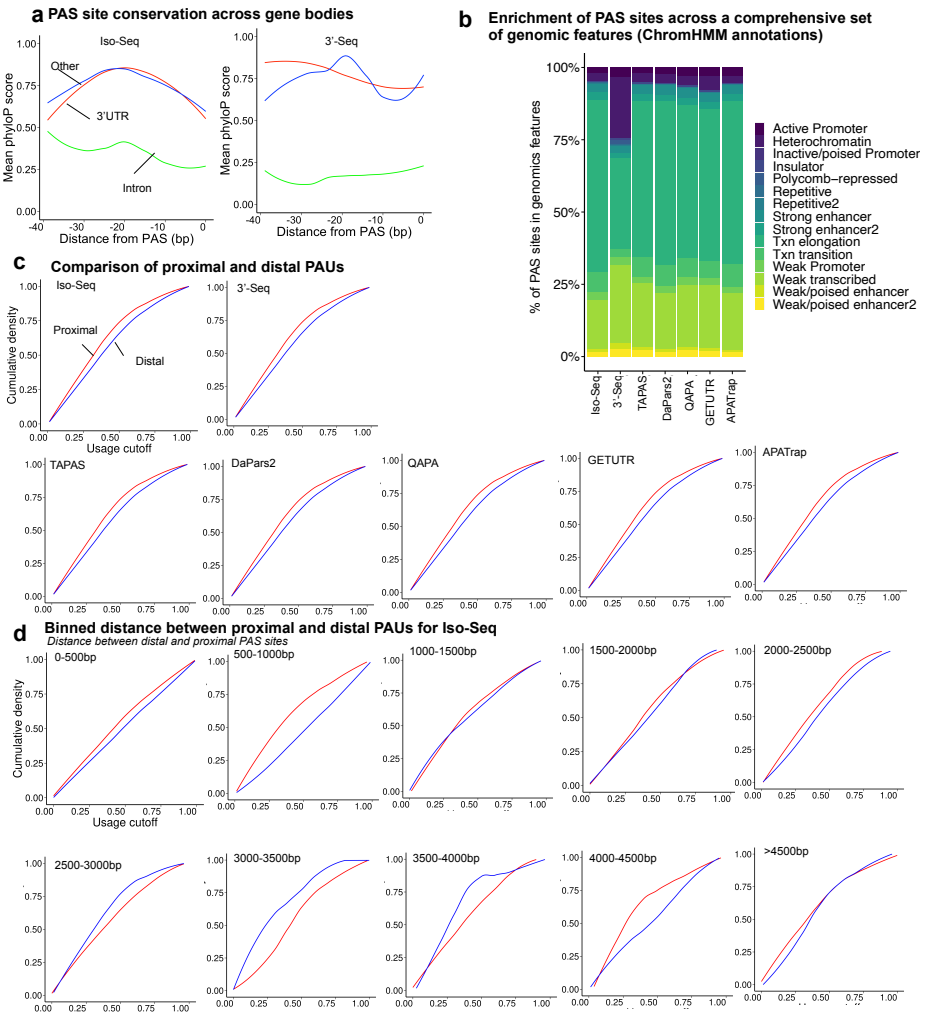


Figure 2.9: **PAS site genic, genomic, and usage features.** (a) Mean phyloP scores of PAS sites, stratified by localization (3' UTRs, introns, or other genic regions labeled as Other). PAS sites localized within 3' UTRs tend to be more conserved than those in introns. (b) Barplots showing the genomic locations of PAS sites using 15 ChromHMM annotations [50]. (c) For every gene within our set of 2,862 genes with $\geq$ 40 Iso-Seq read coverage, we selected PAS sites within 3' UTRs with maximum distance between them and defined the PAS site upstream as the proximal PAS site and the PAS site downstream as the distal PAS site. The proximal and distal PAUs were plotted against varying usage cutoffs. (d) We binned the distal and proximal PAS sites within 3'UTRs among the 2,862 genes with $\geq$ 40 Iso-Seq read coverage by distance (500 bp windows). We observed no evidence of PAU estimation bias for short versus long isoforms using Iso-Seq.

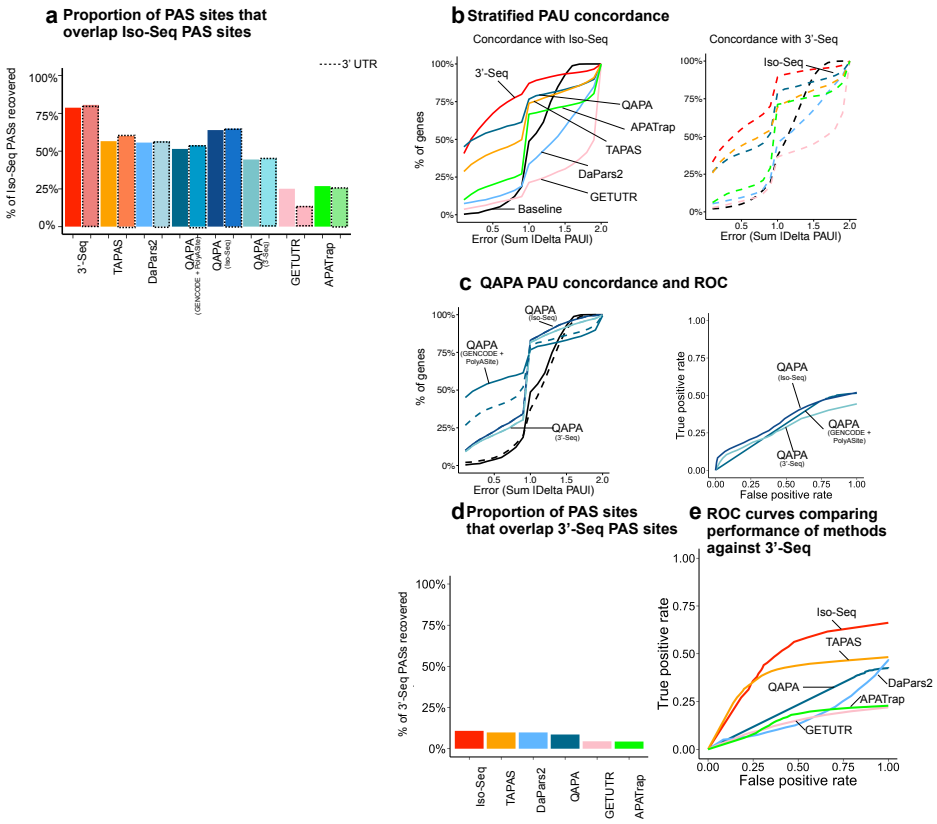


Figure 2.10: **PAS site identification and PAU quantification across Iso-seq and short-read methods, including QAPA run with different PAS site annotations.** (a) Proportion of PAS sites across 2,862 genes identified by short-read sequencing methods, 3'-Seq, QAPA run with three different PAS site annotations, DaPars2, TAPAS, GETUTR, and APATrap that are also identified by Iso-Seq. Bars with dotted lines represent the % of Iso-Seq PAS sites recovered in 3'UTRs only. (b) Comparison of PAU calls across methods. Error(Sum| Delta PAU |) refers to the concordance in calls between two methods, as per Figure 4. The left compares all methods against Iso-Seq. The right compares all methods against 3'-Seq. (c) Error(Sum | Delta PAU |) receiver operating characteristic (ROC) curve, stratified by distinct QAPA runs using different PAS site annotations, including with GENCODE and PolyASite, Iso-Seq, and 3'-Seq. True positives are instances in which Iso-Seq PAS sites with PAUs $> 5\%$ have analogous PAS sites defined by other methods with PAUs $> 5\%$. False positives are PAS sites defined by other methods with PAUs $> 5\%$, but lack analogous PAS sites defined by Iso-Seq with PAUs $>5\%$. (d) Proportion of PAS sites across 2,862 genes identified by Iso-Seq and short-read sequencing methods, QAPA run with three different PAS site annotations, DaPars2, TAPAS, GETUTR, and APATrap that are also identified by 3'-Seq. (e) Receiver operating characteristic (ROC) curves. True positives are instances in which 3'-Seq PAS sites with PAUs $> 5\%$ have analogous PAS sites defined by other methods with PAUs $> 0\%$. False positives are PAS sites defined by other methods with PAUs $> 5\%$, but lack analogous PAS sites defined by Iso-Seq.

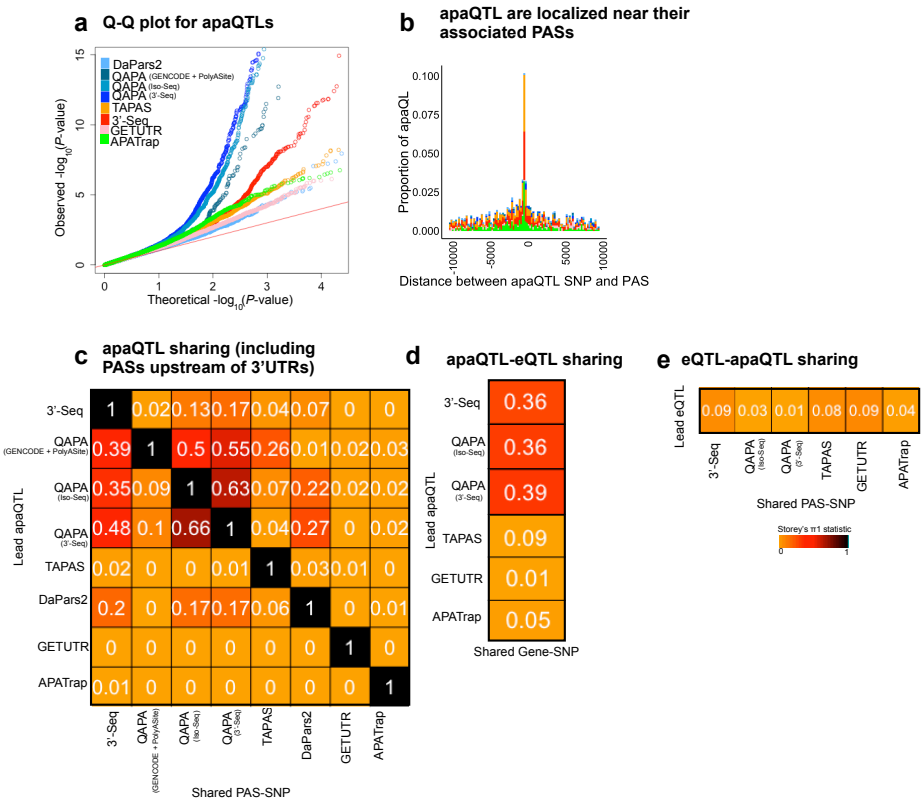


Figure 2.11: **Comparison of apaQTL between APA methods.** (a) QQ-plot showing apaQTL signals, stratified by method. (b) Location of the lead apaQTL SNP relative to its associated PAS site, stratified by method. (c) Quantification of sharing of the impact of genetic variation on APA across sequencing methods using Storey's π_1 statistic. This analysis includes apaQTL linked to PAS sites within and upstream of 3'UTRs. (d) Storey's π_1 statistics quantifying the sharing between the lead apaQTL across the methods and most significant Gene-SNP pair. (e) Storey's π_1 statistics quantifying the sharing between the lead eQTL and most significant PAS-SNP pair per gene (2,864 eQTL, FDR $< 10\%$).

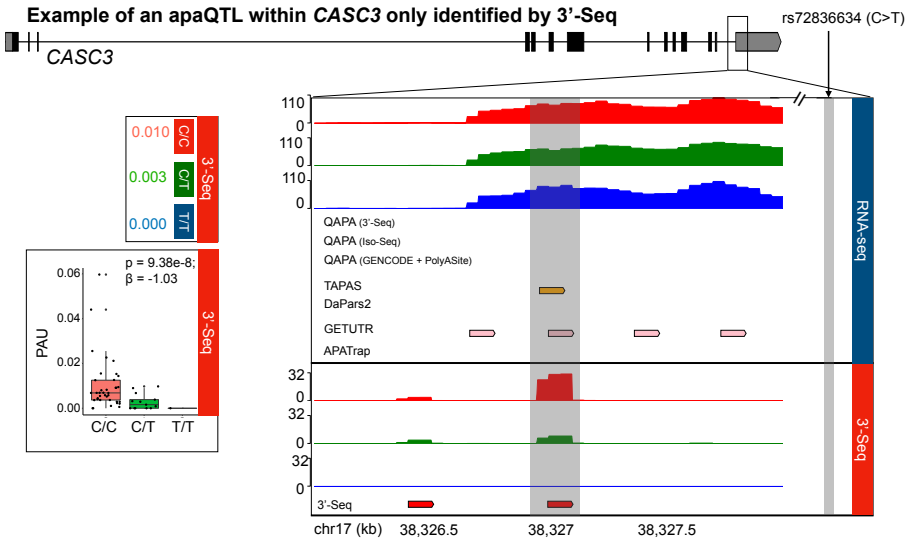


Figure 2.12: **Example of an apaQTL in the CASC3 gene defined by 3'-Seq exclusively.**

2.9 Supplemental tables

Table 2.1: **Iso-Seq.PAS.sites.** Iso-Seq PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.2: **3-Seq.PAS.sites.** 3'-Seq PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.3: **TAPAS.PAS.sites.** TAPAS PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.4: **DaPars2.PAS.sites.** DaPars2 PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.5: **QAPA.GENCODE.PolyASite.PAS.sites.** QAPA (run with GENCODE and PolyASite annotations) PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.6: **QAPA.Iso-Seq.PolyASite.PAS.sites.** QAPA (run with Iso-Seq and PolyASite annotations) PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.7: **QAPA.3-Seq.PolyASite.PAS.sites.** QAPA (run with 3-Seq and PolyASite annotations) PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.8: **GETUTR.PAS.sites.** GETUTR PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

Table 2.9: **APATrap.PAS.sites.** APATrap PAS sites with PAU > 5% (BED format). See supplementary file associated with Shah et al., *Genome Biology*, 2021 [154].

CHAPTER 3

M6A MRNA METHYLATION IS ESSENTIAL FOR OLIGODENDROCYTE MATURATION AND CNS MYELINATION

3.1 Abstract

The molecular mechanisms that govern the maturation of oligodendrocyte lineage cells remain unclear. Emerging studies have shown that N6-methyladenosine (m6A), the most common internal RNA modification of mammalian mRNA, plays a critical role in various developmental processes. Here, we demonstrate that oligodendrocyte lineage progression is accompanied by dynamic changes in m6A modification on numerous transcripts. *In vivo* conditional inactivation of an essential m6A writer component, METTL14, results in decreased oligodendrocyte numbers and CNS hypomyelination, although oligodendrocyte precursor cell (OPC) numbers are normal. *in vitro* Mettl14 ablation disrupts postmitotic oligodendrocyte maturation and has distinct effects on OPC and oligodendrocyte transcriptomes. Moreover, the loss of Mettl14 in oligodendrocyte lineage cells causes aberrant splicing of myriad RNA transcripts, including those that encode the essential paranodal component neurofascin 155 (NF155). Together, our findings indicate that dynamic RNA methylation plays an important regulatory role in oligodendrocyte development and CNS myelination.

3.2 Introduction

Oligodendrocytes are glial cells that are responsible for myelination in the central nervous system (CNS). Myelin is a multilayered membrane sheath that insulates axons and is important for a myriad of CNS functions, including providing metabolic support to axons and allowing for rapid and efficient propagation of electrical signals. Recent studies suggest that

myelin sheath plays a critical and active role in CNS functions, including being implicated in sensory experience, aging, memory, and motor skill learning [25]. Defects in myelination are associated with developmental disorders and neurodegenerative diseases, such as multiple sclerosis (MS) [17].

During development, oligodendrocyte progenitor cells (OPCs) arise from neuroepithelial cells in the ventricular zone in mice at embryonic day 12.5 (E12.5) and in humans during gestational week 6.5 (E45) [16]. By E15 in mice, the OPCs proliferate and migrate to their final destination, where they terminally differentiate into mature, myelinating oligodendrocytes (Bergles and Richardson, 2015). Several factors have been identified as critical regulators of oligodendrocyte development. For example, OPCs express several transcription factors, including OLIG2, SOX10, NKX2.2, ZFP24, and MYRF, as they migrate and mature into oligodendrocytes [49] [120]. Several signaling pathways are also critical for oligodendrocyte lineage progression and maturation [66].

Epigenetic mechanisms, including DNA methylation, histone modification, ATP-dependent chromatin remodeling, and gene silencing by non-coding RNAs, including long noncoding RNAs and microRNAs, have also been implicated [98] [113] [123]. For example, in MS patients, the DNMT family is upregulated and TET families are downregulated in the hippocampus [28]. A study conducted on MS brains highlighted that oligodendrocyte survival genes were hypermethylated and lowly expressed as compared to those in control brains [78]. In comparison, proteolytic processing genes were hypomethylated and highly expressed, revealing that DNA methylation changes are possible contributors of MS occurrence. In contrast to the study of reversible chemical modification of DNA, the study of RNA modifications as regulators of gene expression has only recently been initiated [56] [192], and its role in oligodendrocyte development remains elusive.

N6-methyladenosine (m6A) is the most abundant internal RNA modification, deposited on both polyadenylated mRNA and non-coding RNA, occurring at a frequency of 1-2% per

nucleotide [85] Levels of m6A are regulated by m6A methyltransferase complexes, or reader proteins, m6A demethylases, or writer proteins [56]. Methyltransferase-like 3 (METTL3) was discovered to be the core catalytic subunit of the m6A mRNA methyltransferase complex [18] [43] [67]. Other subunits that were subsequently identified include methyltransferase-like 14 (METTL14), which facilitates binding of the methyltransferase complex to the RNA [109], and Wilms tumor 1-associating protein (WTAP), which recruits the METTL3-14 heterodimer [109] [139]. Two m6A erasers, ALKBH5 and FTO, have been identified [180] [81]. Most reader proteins contain a YT521-B homology (YTH) domain, which binds RNA in an m6A-dependent manner [47]. RNA metabolism, including stability, translation, localization, and splicing, can be regulated by m6A [147].

Moreover, in contrast to DNA and protein methylation, m6A methylation has the potential to have a very rapid influence on transcriptome changes during cell state transitions, such as cell differentiation and development [55] [198]. A recent study in neural stem cells revealed that conditional inactivation *METTL14* disrupts cortical neurogenesis [191], thus highlighting the critical role that m6A plays in the CNS. Nevertheless, the impact of m6A on oligodendrocyte lineage regulation has remained unclear. Therefore, in this study, we and our collaborators sought to better understand the role that m6A plays in oligodendrocyte lineage progression by conditionally inactivating *Mettl14* in OPCs using a Mettl14 conditional (floxed) mouse line in combination with oligodendrocyte Cre driver lines. *In vitro*, OPCs lacking Mettl14 did not properly differentiate into mature oligodendrocytes, suggesting that m6A plays a critical role in oligodendrocyte differentiation. RNA sequencing (RNA-seq) and SMART-Seq m6A-seq revealed that OPC and oligodendrocyte transcripts encoding transcription factors, DNA epigenetic regulators, and signaling pathways that are critical for oligodendrocyte lineage progression were m6A marked and differentially affected by the Mettl14 deletion. We also found pervasive aberrant mRNA splicing in the Mettl14-deleted OPCs and oligodendrocytes. Importantly, we discovered that the critical paranode

component NF155 is differentially spliced and significantly disrupted during myelination in the Mettl14-ablated mutants.

3.3 Methods

For a full list of methods, please see Xu et al., *Neuron*, 2020 [188]. Methods included below are methods specific to analyses presented in this dissertation.

3.3.1 Total protein and RNA isolation

Protein from cells and snap frozen half-brain was isolated as previously described [31] Protein concentration was determined using a BCA Protein Assay Kit (Thermo Fisher Scientific, cat# 23255). RNA from cells and snap frozen half-brain was isolated as previously described [179]. RNA quality was confirmed by 2100 Bioanalyzer using a model 6000 Nano kit (Agilent technologies, cat# 5067-1511). Samples with an RNA integrity number ≥ 8 were used.

3.3.2 RNA-seq and analysis

Bulk RNA-seq was performed on RNA isolated from cultured OPCs and oligodendrocytes as previously described [3]. Libraries were prepared and sequenced using the Illumina HiSeq 4000 at the University of Chicago Genomics Core facility. Reads were mapped using both STAR v2.6.1a and Kallisto v.0.44.0 using bowtie 2 aligner [19] [46]. Mapped reads were further analyzed with the Bioconductor suite v3.7 by the University of Illinois at Chicago Bioinformatics Core facility [77]. Q-values were determined as false discovery rate adjusted p-values using the method previously described [15]. Results were compared with the m6A-SMART-Seq analysis and visualized in R v.3.5.1 using the plot.ly, ggplot2, and venn.diagram packages. Values for expression, fold change and statistical significance were adapted for visualization using a log2 transformation. Thresholds were set to 1, +/-1.5, and 0.001,

respectively.

3.3.3 m6A-SMART-seq and analysis

mRNA from total RNA of OPCs and oligodendrocytes was purified with Dynabeads Oligo (dT) (Thermo Fisher Scientific). The purified mRNA was then processed for m6A-SMART-seq and analyzed as previously described [182]. Z scores were calculated for each m6A mark and filtered with a threshold value of 0.

3.3.4 Differential alternative splicing analysis

Differential splicing analysis was performed between OPCs versus OPCs lacking Mettl14 and oligodendrocytes versus oligodendrocytes lacking Mettl14. In brief, exon-exon junctions from mapped RNA-seq reads, which are representative of introns that are removed from pre-mRNA, were extracted. Next, alternatively excised introns, which are comprised of two more overlapping introns (e.g., introns that share a splice site), were clustered together. Finally, differential intron excision events across conditions were tested using LeafCutter [103].

3.3.5 Data Availability

The sequencing data have been deposited to the National Center for Biotechnology Information Gene Expression Omnibus (GEO) database under accession number: GSE124244.

3.4 Results

3.4.1 Oligodendrocyte lineage progression is accompanied by changes in m6A modification on numerous transcripts

RNA modification by the m6A mark has emerged as an important mechanism to regulate gene expression during cell lineage development [55]. To profile m6A and its impact on gene expression during oligodendrocyte lineage progression, our collaborators in the Popko lab collected m6A-seq and RNA-seq data from purified OPCs and oligodendrocytes. Their manuscript provides more information on the purification process [188]. Bulk m6A-seq was difficult to perform given the insufficient mRNA yield. Therefore, our collaborators used a SMART2 single cell RNA-seq method, m6A-SMART-seq, for sensitive full-length m6A profiling in single cells [134] [182]. They detected 3,554 transcripts bearing m6A in OPCs and 2,606 transcripts bearing m6A in oligodendrocytes (Xu et al., *Neuron*, 2020). Gene ontology analyses indicated that these m6A marked transcripts are important functions for cell development in both OPCs and oligodendrocytes (Xu et al., *Neuron*, 2020 [188]) (log2 (CPM) ≥ 1, Z score ≥ 0). Of the 11,502 genes expressed in both OPCs and oligodendrocytes, 23 transcripts bore m6A in both OPCs and oligodendrocytes, 2,806 transcripts bore the m6A in OPCs exclusively, and 1,626 transcripts bore m6A in oligodendrocytes exclusively (Xu et al., *Neuron*, 2020 [188]).

3.4.2 Mettl14 ablation differentially alters OPC and oligodendrocyte transcriptomes

Our collaborators generated mouse lines in which METTL14, an essential m6A writer component, was conditionally inactivated, in both developing oligodendrocyte lineage cells (Mettl14fl/fl;Olig2-Cre) and postmitotic, maturing oligodendrocytes (Mettl14fl/fl;CNP-Cre). The manuscript outlines additional details on the mouse crosses [188].

To determine the effects of m6A on gene expression during oligodendrocyte lineage progression, our collaborators collected RNA-seq data from purified OPCs and oligodendrocytes with both purified OPCs and cultured mature oligodendrocytes from Mettl14fl/fl;Olig2-Cre control and mutant mice. See their manuscript for more information on the purification process [188].

They detected 11,809 transcripts present in the OPCs, of which 586 were significantly upregulated in mutant cells and 177 were significantly downregulated in mutant cells. Among the 12,542 transcripts present in mature oligodendrocytes, 1,388 transcripts were significantly upregulated and 1,247 were significantly downregulated in the mutant cells (Xu et al., *Neuron*, 2020 [188]). Importantly, the significantly downregulated transcripts are normally expressed in myelinating oligodendrocytes and encode myelin-protein expression factors, such as Mbp, Mog, Mag, Plp1, and Cnp.

Upon comparing the m6A-seq data and RNA-seq data, our collaborators found that of the 3,554 m6A marked OPC transcripts, 46 were significantly downregulated in mutant cells and 108 were significantly upregulated in mutant cells (Xu et al., *Neuron*, 2020 [188], Figure 6C). Of 2,606 m6A marked oligodendrocyte transcripts, 221 were significantly downregulated in mutant cells and 217 were significantly upregulated in mutant cells (Xu et al., *Neuron*, 2020 [188]). Gene ontology analysis indicated many important functions such as glia cell development in OPCs (Xu et al., *Neuron*, 2020 [188]) and myelination in oligodendrocytes (Xu et al., *Neuron*, 2020 [188]). These results suggest that the m6A mark differentially regulates the OPC and oligodendrocyte transcriptomes.

3.4.3 Mettl14 regulates transcripts that encode transcription factors that are critical for oligodendrocyte lineage progression

In order to elucidate how the m6A mark regulates oligodendrocyte lineage development, our collaborators compared m6A-seq and RNA-seq datasets across OPCs and oligodendrocytes

upon *METTL14* ablation to identify transcripts marked by m6A that encode transcription factors known to be involved in oligodendrocyte lineage progression. For example, it is known that transcription factors, such as Nkx-2.2, Olig1, Olig2, Sox10, Myrf, and ZFP24, are the major determinants of oligodendrocyte differentiation and myelination [49] [120].

Our collaborators identified that transcripts encoding *Hey1, Klf19, Sox2, Sox5, Srebf1 ,Tcf19, Zeb2* were marked by m6A in OPCs but not in oligodendrocytes. Notably, misregulation of SOX2 can impair OL differentiation [74]. Moreover, sterol regulatory element binding proteins (SREBPs), including Srebf1, control fatty acid and cholesterol metabolism, and recent studies suggest that SREBPs enable myelin lipid synthesis in oligodendrocytes and are controlled by the mammalian target of rapamycin (mTOR) pathway [54]. Therefore, impaired SREBP would result in impaired myelination. ZEB2 plays a role in OPC differentiation and reduces the number of OLs upon impairment [181].

They also discovered that transcripts encoding *Hes1, Nkx6.2, Olig2 and Yy1* were marked by m6A in oligodendrocytes but not in OPCs. Notably, upon impairment, YY1 arrests OL differentiation as they exit the cell cycle [68]. Phenotypically, this is presented as defective myelination, ataxia, and tremor. At the molecular level, YY1 normally recruits histone deacetylase-1 to the promoters of of transcriptional inhibitors of myelin genes during oligodendrocyte differentiation, thus repressing these inhibitors, however, upon impairment, these transcriptional inhibitors are expressed and these myelin genes remain downregulated.

3.4.4 Mettl14 regulates transcripts that encode histone acetyltransferases, methyltransferases, lysine demethylases that are critical for oligodendrocyte lineage progression

Epigenetic regulation drives oligodendrocyte lineage progression and myelination. Previous studies have demonstrated DNA epigenetic regulation mechanisms, including histone modifications, are important for oligodendrocyte lineage progression [90]. These sequencing anal-

yses were concordant with previous studies in that they similarly revealed that transcripts encoding histone modification regulators bear m6A and were significantly differentially expressed in mutants. Our collaborators discovered transcripts of histone writers, such as histone acetyltransferases (HATs) Hat1, histone methyltransferases (HMTs) Smyd2, Prdm2, Setdb1, Suv39h1, Ash1l, Dot1l, histone erasers, such as histone deacetylases (HDACs) Hdac3, Hdac7, Hdac8, Hdac9, and lysine demethylases (KDMs) Kdm2b, Kdm5c, Kdm3b , Kdm4a, Kdm4c, Kdm6a.

Normally, these transcripts encode proteins with important regulator functions in oligodendrocyte development [70]. Interestingly, recent evidence has also demonstrated an essential functional role of m6A on RNA modification on histone modifications in regulating embryonic neural stem cell self-renewal in the CNS [182]. Thus, our findings suggest a possible link between m6A RNA modification and histone modifications in the regulation of oligodendrocyte lineage development.

3.4.5 Mettl14 regulates transcripts that encode key signaling pathway molecules that are critically involved in oligodendrocyte lineage progression

Transcripts that were significantly altered by Mettl14 ablation encoded molecules involved in signaling pathways and their signaling molecules. These included bone morphogenetic proteins (BMPs), ERK/MAPK, fibroblast growth factor families (FGFs), Notch/Delta, Sonic hedgehog (Shh), and Wnt signaling pathways in OPCs, and P13K/AKT/mTOR, BMPs, ERK/MAPK, insulin-like growth factor-1 (IGF-1), Notch/Delta, Shh, and Wnt signaling pathways in oligodendrocytes. These pathways have previously been associated with oligodendrocyte development and myelination regulation [66]. Our collaborators observed that many transcripts encoding critical components of these signaling pathways harbored the m6A mark, suggesting that m6A may regulate these signaling pathways to promote oligo-

dendrocyte lineage progression.

3.4.6 Mettl14's possible mechanisms of action in oligodendrocyte lineage cells

Various studies have demonstrated that m6A influences various aspects of mRNA metabolism, including stability, translation, localization, and splicing [147] [177] [186] [199] [203]. The influence of m6A on these processes is mediated by RNA-binding proteins. For example, the YTH domain-containing proteins were first identified as m6A readers, which bind RNA in an m6A-dependent manner [47] [193] [205]. Additionally, m6A recruits RNA-binding proteins indirectly to execute functions, such as by altering RNA folding structures to influence RNA-binding protein access to the transcript and binding efficiency [110].

Given this, we aimed to explore potential mechanisms of action of the m6A mark in regulating oligodendrocyte lineage cell development and function, in addition to the direct effects on gene expression, which we described previously. Previous studies have demonstrated that transcripts bearing m6A have reduced stability [182] [191] [194]. Indeed, in our animal models lacking transcripts bearing m6A given ablation of Mettl14, we and our collaborators did observe higher expression levels of *Hey1, Sox5, Hac9*, and *Setd1b* in OPCs and *Hes1, Nkx6.2, Yy1, Hdac7, Kdn4c*, and *Hat1* in oligodendrocytes. However, a substantial number of transcripts did not demonstrate differences in stability, suggesting that indirect mechanisms might contribute to gene expression changes.

Previous studies have demonstrated that transcripts bearing m6A have increased translational efficiency [36] [157]. Upon comparing transcriptional and translational levels of m6A marked transcripts that encode proteins critical for oligodendrocyte development – Myrf, Olig2, Mbp, and Mag – we and our collaborators observed decreased levels of these proteins [21] [204]. Interestingly, these observed reductions correlated with the reductions in m6A found in the oligodendrocyte transcriptome, suggesting that translational regulation may

not be a key feature of m6A gene regulation in oligodendrocyte lineage cells.

3.4.7 Mettl14 ablation does not disrupt Mbp transport

Different isoforms of myelin proteins are generated by alternative splicing, and the different isoforms ensure precise oligodendrocyte lineage progression [196] [200]. Previous studies have demonstrated that m6A plays a critical role in regulating mRNA splicing [200]. In order to investigate the potential role of m6A in regulating differential splicing during oligodendrocyte development specifically, we used LeafCutter [103] to identify altered splicing events in OPC and oligodendrocyte transcriptomes. LeafCutter identifies alternatively excised intron clusters and compares differentially excised intron levels between controls and mutants. Differential splicing is measured by changes in the percent spliced in (or delta, dPSI) [103].

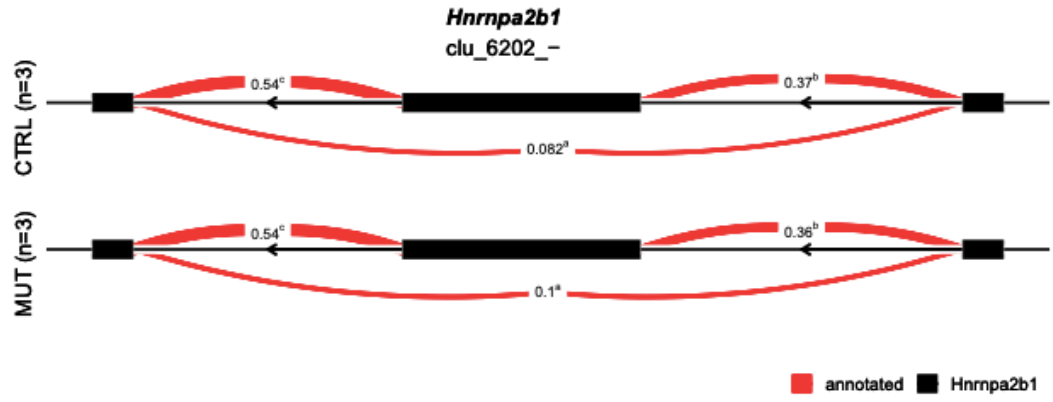
Transcripts with highest dPSI level

Oligodendrocytes				
Gene	dPSI	log_fc	log_cpm	qval
Mettl14	0.93	-1.52	5.01	1.58E-85
Fhl3	0.80	-1.43	3.46	1.55E-54
Wdr91	0.50	-0.16	5.29	1.09E-01
Pms2	0.49	1.37	2.69	5.32E-29
Sh3bp2	0.47	-1.15	3.92	6.42E-34
Atp9a	0.46	-0.51	7.77	1.10E-12
Nfasc	0.45	-0.87	10.81	1.50E-29
Aasdh	0.44	1.00	4.48	2.21E-31
Lpgat1	0.43	-0.51	6.22	8.63E-11
Mapk8ip1	0.43	-1.75	8.40	4.44E-112

OPCs				
Gene	dPSI	log_fc	log_cpm	qval
Hexdc	0.61	0.31	3.87	4.69E-02
Ptprz1	0.56	-0.52	11.01	3.74E-10
Gm44503,Ccpg1	0.56	-0.22	2.20	4.04E-01
Mettl14	0.53	-1.18	5.30	3.07E-40
Heca	0.50	0.93	6.05	2.50E-20
D11Wsu47e	0.47	0.69	4.28	1.05E-08
Rfesd	0.42	-0.32	3.31	7.68E-02
Dixdc1	0.41	0.73	4.27	5.17E-11
Ppcdc	0.40	-0.68	3.38	1.17E-06
Znrf1	0.39	0.31	6.10	1.35E-03

Table 3.1: **Top 10 dPSI aberrantly spliced transcripts (oligodendrocytes and OPCs).** LeafCutter analysis showing aberrantly spliced transcripts in Mettl14fl/fl;Olig2-Cre mutant oligodendrocytes and OPCs. Listed are the top 10 transcripts that have the highest dPSI values, with their chromosome intron locations and Log2 effect sizes. (dPSI: changes/or delta in the percent spliced in; n=3)

We observed that the m6A mark had a pervasive impact on OPC and oligodendrocyte mRNA alternative splicing (Table 3.1). Interestingly, we observed that m6A impacted alternative splicing of the family of RNA binding proteins, heterogeneous nuclear ribonucleoprotein in oligodendrocytes (hnRNPs). Previous studies have demonstrated that hnRNPs are m6A readers [147]. Moreover, additional studies have demonstrated that hnRNPs are regulators of splicing [86] Therefore, this suggests that m6A might be able to indirectly impact the alternative splicing of transcripts through different hnRNP isoforms. For example, an isoform of hnRNPA2, hnRNPA2B1, which is expressed in oligodendrocyte lineage cells, is an m6A reader [6]. Upon recognition of transcripts bearing m6A in the nucleus, hnRNPA2B1 enables proper splicing of these transcripts [6].



	chr	start	end	verdict	dPSI
a	chr6	51464211	51465179	annotated	0.019
b	chr6	51464547	51465179	annotated	−0.012
c	chr6	51464211	51464428	annotated	−0.006

Figure 3.1: **Mettl14 ablation differentially alters hnRNPA2 alternative splicing.** Schematic view of differentially spliced sites in the hnRNPA2B1 gene in control versus Mettl14fl/fl;Olig2-Cre mutant day5 oligodendrocytes (N=3). Each cluster (i.e., abbreviated as clu) represents a group of introns that display alternative excision events. Specifically, these are introns that share a donor site (canonical 50 splice site, AT) or acceptor site (canonical 30 splice site, GA). Red curves represent cases with more splicing events in the mutants ($p \leq 0.05$). This cluster highlights exon 10 (middle) of hnRNPA2B1.

The m6A mark has also been shown to play a role in intracellular mRNA transport [147], and this could be mediated indirectly through m6A and alternative splicing. Myelin basic protein (Mbp) is critical myelin protein that is translated locally in the myelin compartment [32] [24]. MBP requires active transport of its mRNA from the nucleus to the cytoplasm,

which is mediated by the hnRNPA2 splicing isoform, hnRNPA2B1 [73] [125]. Specifically, hnRNPA2B1 recognizes the A2 response element (A2RE), a cis-acting signal present in certain trafficked mRNAs, including the mRNA that encodes Mbp [64]. We aimed to assess if m6A might have a role in regulating Mbp mRNA transport in oligodendrocytes, as mediated by hnRNPA2B1. We observed differential splicing of hnRNPA2B1 between controls and Mettl14 ablated oligodendrocyte mutants (Figure 3.1), although these differences were not statistically significant. For example, exon 10 of hnRNPA2B1 was more frequently spliced out in mutants, and this exon plays a critical role in hnRNPA2B1's recognition of A2RE signals to transport mRNAs [64]. Given this, we hypothesized that Mbp transport to the cytoplasm might be disrupted in mutants.

To assess if Mbp transport had been disrupted, our collaborators used RNAscope, which is an *in situ* hybridization assay for detection of target RNA within intact cells, to determine the distribution of Mbp distribution in oligodendrocytes of the corpus callosum [175]. Interestingly, in mouse models with Mettl14 ablated, our collaborators observed reduced levels of Mbp mRNA overall but did not observe that their distribution had been altered as compared to controls. These results suggest that the while the absence m6A had altered the splicing of hnRNPA2, it had not disrupted sub-cellular transport of the Mbp mRNA in the myelin compartment.

3.4.8 Mettl14 ablation differentially alters Nfasc155 alternative splicing

At the global level, we observed that 1,372 splicing events across 364 genes in OPCs and 1,930 splicing events across 485 genes in oligodendrocytes that were differentially spliced upon Mettl14 ablation ($q < 0.01$). A number of these significantly differentially alternative spliced transcripts have previously been shown to encode proteins with important functions in the myelinating process. Notably, Nfasc had one of the most significantly altered isoforms and bore one of the highest differential dPSI level in the oligodendrocyte transcriptome

(Table 3.1).

Nfasc is essential in the establishment and maintenance of node of Ranvier domains [75] [138] [155]; [164] [206]. The mouse Nfasc gene contains 39 exons, and inclusion or exclusion of different exons results in transcripts that encode functionally distinct isoforms [160]. For example, Nfasc186 is expressed by neurons and is critical for node assembly. In contrast, Nfasc155 is expressed by the myelinating cells and is critical for the stability of the paranodal domain [75].

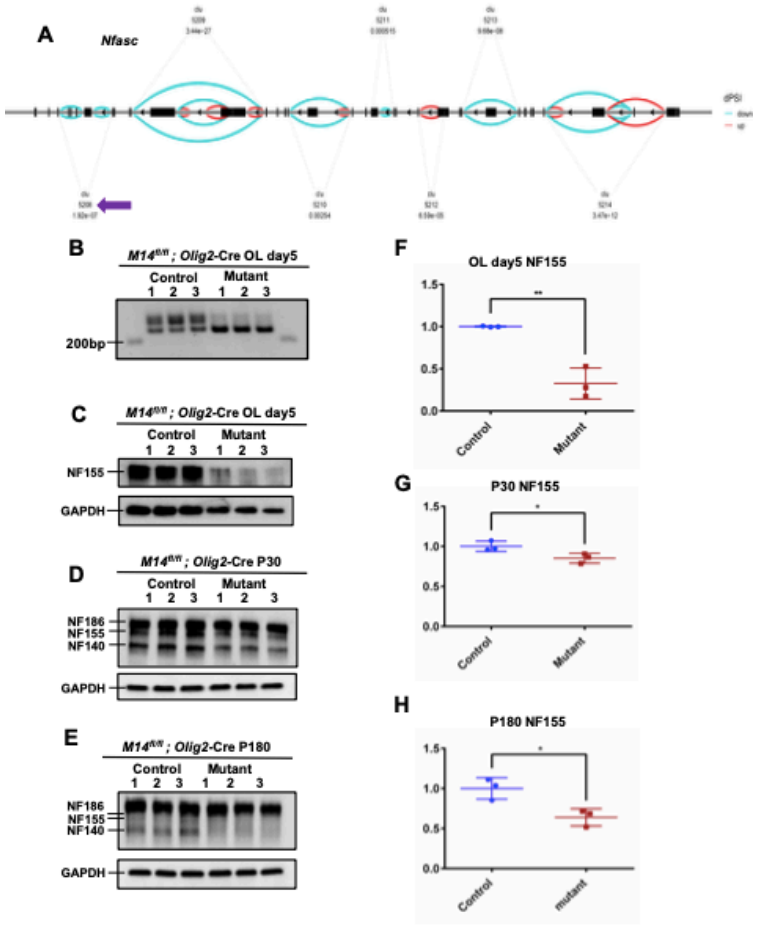


Figure 3.2: **Mettl14 ablation differentially alters Nfasc155 alternative splicing.** (A) Schematic view of differentially spliced sites in the Nfasc gene in control versus Mettl14fl/fl;Olig2-Cre mutant day5 oligodendrocytes. The 39 Nfasc exons are labeled above the exons. Each cluster (i.e., abbreviated as "clu X") represents a group of introns that display alternative excision events. Specifically, these are introns that share a donor site (canonical 50 splice site, AT) or acceptor site (canonical 30 splice site, GA). Blue curves represent cases that have fewer splicing events in the mutants, while the red curves represent cases with more splicing events in the mutants ($p \leq 0.05$).

Figure 3.2 *(cont'd)*: (B) Differentially spliced Nfasc isoforms were detected by RT-PCR and agarose gel electrophoresis in the Mettl14fl/fl;Olig2-Cre day5 oligodendrocyte mutants (218 kb). Primers used: Forward, ACTGGGAAAGCAGATGGTGG; Reverse, ACATGAGCCCGATGAACCAG. (C-E) Western blot results of NFASC (P30, P180). (F-H) Quantification of NFASC155 expression(P30, P180). NFASC155 expression level was normalized to GAPDH expression level. NFASC155 had significant reduction in both P30 and P180 Mettl14fl/fl;Olig2-Cre mutants.

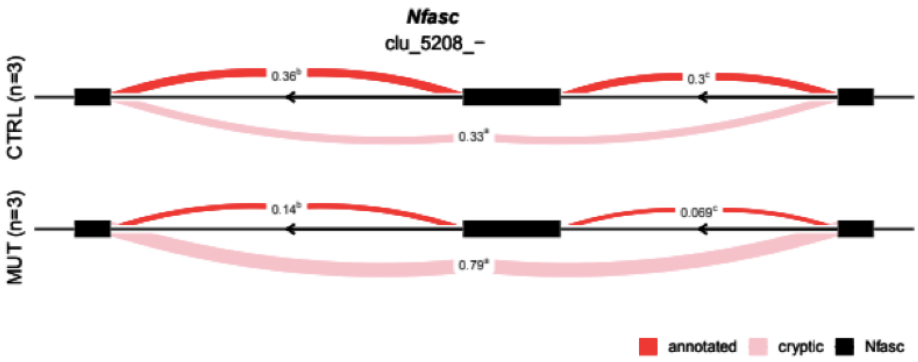


Figure 3.3: **Mettl14 ablation differentially alters Nfasc155 (exon 25)alternative splicing.** The magnified window shows the sample cluster (clu 5208) that we examined for the presence of aberrant spliced isoforms in the mutants in (Figure 3.2. This cluster highlights exon 25 of Nfasc155.

The disruption of Nfasc isoform distribution can result in pathological changes in myelinated axons [75] [138] [164]. Notably, we observed that exon 25 in cluster 5208, a cluster of introns important for Nfasc155 isoform, was spliced out in Mettl14 ablated animals as compared to controls (Figure 3.3A). Our collaborators performed RT-PCR to confirm the abundance of differentially spliced Nfasc isoforms (Figure 3.2) in purified oligodendrocyte mRNA from mutants and controls. Our collaborators performed additional experiments to examine Nfasc155 expression levels at different developmental stages. Western blot results revealed that 1 month old (P30) and adult (P180) animals showed significantly decreased Nfasc155 levels as compared to controls (Figure 3.2).

Interestingly, exon 25, which was spliced out in mutants, is what encodes the extra fibronectin type 3 (FNIII) repeat that is unique to the Nfasc155 isoform and is missing from

Nfasc186. FNIII is critical for enabling paranodal axoglial contacts and junction formation [161]. To assess if Mettl14 ablation resulted in aberrant node and paranodes, our collaborators performed immunohistochemistry with antibodies against voltage gated sodium channel (NaCh) and Caspr to identify the nodal and paranodal domains, respectively. They observed that Mettl14 ablation resulted in in aberrant node and paranodes, both in terms of numbers and morphology in the adult (P180) animals (Fig 2.8 E-G). This suggests that loss of m6A results in pathological changes at the nodes of Ranvier. Moreover, the aberrant nodes of Ranvier observed in the Mettl14-deficient mice were reminiscent of Nfasc155-deficient mice [138]. These results indicate that m6A regulates Nfasc155 splicing and plays a role in establishing and maintaining normal function of critical axonal-oligodendrocyte interactions.

We and our collaborators did not assess the relationship between m6A and the RNA binding protein quaking (QKI) in the context of the aberrant nodes of Ranvier observed in the Mettl14-deficient mice. QKI is involved in paranodal axoglial junction formation; specifically, it normally promotes the inclusion of exon 25 through its binding of Nfasc RNA [39]. Lack of m6A marking the Nfasc transcript might have occluded QKI from binding, either directly or indirectly, which would need to validated by methods that allow one to map protein-RNA interactions, such as cross-linking and immunoprecipitation (CLIP) [63]. It would also be useful to confirm that aberrant nodes of Ranvier observed in the Mettl14-deficient and Nfasc155-deficient mice are similar to those observed in QKI-deficient mice.

3.5 Discussion

RNA modifications have recently emerged as critical post-transcriptional regulatory mechanism to modulate gene expression [55]. m6A is the most abundant RNA modification found in eukaryotes [192]. Our study demonstrates that m6A RNA modification is essential for normal oligodendrocyte maturation and CNS myelination. We highlight that the m6A plays an important role in regulating various aspects of gene expression in oligodendrocyte lineage

cells, with the most profound effects on alternative splicing.

We and our collaborators observed dynamic changes of m6A marked status in the transcripts that are expressed in both OPC and maturation stages, which suggests that m6A RNA methylation accompanies differentiation and maturation. To understand the mechanisms of m6A's role in regulating oligodendrocyte development, we compared the transcriptomes of both purified OPCs and mature oligodendrocyte mutants and controls. We observed dramatic differences of differentially expressed transcripts in these two lineage stages and determined that m6A regulates the transcriptomes in OPCs and mature oligodendrocytes. In addition, we discovered many oligodendrocyte lineage regulators, including as signaling pathways and histone modifiers.

Notably, we examined m6A's impact on alternative splicing during oligodendrocyte maturation and identified many aberrantly transcribed transcripts in Mettl14 ablated animals. Nfasc155, a glia isoform that encodes essential protein in the establishment and maintenance of node of Ranvier domains, was among the transcripts identified. We experimentally validated Nfasc155's transcriptional and translational levels and consequences of its aberrant splicing at various developmental stages in mutant animals. Our results confirmed that m6A plays an important role in establishing and maintaining normal function of critical axonal-oligodendrocyte interactions.

3.6 Acknowledgement of work performed

I would like to acknowledge the individuals who contributed to this work. The analyses presented in this chapter (Figure 3.2), (Figure 3.3) were conceived, generated, and visualized by Ankeeta Shah. All experimentation was performed by individuals in Brian Popko's group and Chuan He's group. The remainder of this chapter is adapted, summarized, and paraphrased from the published work (Xu et al., Neuron, 2020 [188]).

CHAPTER 4

GENETIC CONTROL OF NOISY SPLICING UNDERLIES UNEXPLAINED ASSOCIATIONS TO COMPLEX TRAITS

4.1 Abstract

RNA splicing is an error-prone process, which can generate a number of spurious transcripts that are unlikely to be functional. Specific mis-splicing events are rare and are broadly ignored in data analysis because they are assumed to be inconsequential. However, we demonstrate that the amount of error-prone splicing, or noisy splicing, is largely determined by genetic sequence and is optimized for highly expressed genes. By applying an updated version of LeafCutter to the GEUVADIS Consortium lymphoblastoid cell lines, we identified 3,269 splicing fidelity QTL (sfQTL) that correlate with noisy splicing levels at 10% false discovery rate (FDR). Notably, we discovered that sfQTLs colocalize with signals from several genome-wide association studies (GWAS), including inflammatory bowel diseases (IBD). Thus, sfQTLs aid in the interpretation of functional variants beyond standard molecular QTL and can be readily identified by applying an updated version of LeafCutter on existing RNA-seq datasets.

4.2 Introduction

The mammalian transcriptome harbors hundreds of thousands of mRNA transcripts, many of which serve critical functions. Indeed, faithful and specific pre-mRNA splicing is critical for the accurate expression of genes and the production of protein isoforms that have been associated with a myriad of biological functions. The fidelity of splicing is achieved by combinatorial recognition of splice sites and RNA binding protein (RBP) motifs by the multimeric protein complex known as the spliceosome and auxiliary splicing regulators, re-

spectively [152]. Achieving high specificity is an important yet daunting task for the cell as about one-third of the human genome is comprised of introns [137], which is a large sequence space containing many sequence elements similar to consensus splice sites and RBP motifs. Moreover, while many sequences in the mammalian genome match the consensus, they might not be recognized as real splice sites, for example, and how the spliceosome makes a choice amongst splice sites during kinetic competition between splicing and transcription is poorly understood [149]. Tight quality control mechanisms must be in place to prevent the production of a large number of aberrantly spliced transcripts. While the fidelity of splicing is a tightly regulated process, mediated by a number of trans-acting protein factors [152], splicing errors may still occur, albeit less than 1% of the time per intron [137]. Error-prone splicing gives rise to a number of low-abundance, non-functional, and unconserved transcripts, which we consider to be *noisy* splicing events [137].

A previous study demonstrated that mRNA splicing is a primary mechanism that links genetic variation to disease [104]. Genetic variants can impact RNA splicing by disrupting recognition of canonical splice sites or splicing regulatory elements, resulting in aberrant mRNA transcripts, which can cause a large array of human diseases [151]. For example, patients with monotonic dystrophy type 1 have an expanded CUG repeat in the 3'UTR of the DM protein kinase gene, which results in the sequestration of splicing regulators in the muscle-blind protein family, disrupting a number of muscle-blind-dependent splicing events [83] [107]. Moreover, several other studies have underscored this point by identifying splicing quantitative trait loci (sQTL), which are genetic variants associated with splicing, in linkage disequilibrium with genome-wide association (GWAS) hits for a number of diseases, including Type 2 diabetes, Alzheimer's disease, and schizophrenia [51, 143, 163]. This suggests that SNPs affecting splicing have the potential to be tagging causal variants underlying a substantial number of GWAS hits.

One limitation of sQTL analysis is that it is often difficult to interpret the direction of

effect for sQTL because two splicing events are often negatively correlated. Thus, the field generally believes that a SNP that promotes the production of transcripts from one functional splicing isoform also reduces the production of transcripts from another functional splicing isoform. This is traditionally what we think of as alternative splicing. However, we believe that a substantial fraction of sQTL act through distinct mechanisms. For example, it is also possible that a SNP can disrupt the recognition of a canonical splice site, either by directly altering splice sites or by indirectly altering splicing regulatory sequences, thus leading to an opportunity for a number of weaker, cryptic splice sites to be chosen, rather than a different canonical splice site getting chosen. Moreover, a SNP might directly enhance the recognition of a cryptic splice site such that it can actively compete for recognition with the canonical splice site. In both of these situations, one will observe an increase in the usage of cryptic splice sties, thus resulting in increased production of non-functional transcripts, or we simply see a reduction in the functional transcript being produced. This is possible as long introns in human transcripts provide ample sequence space for mutations to arise that result in the creation of new and sometimes cryptic splice sites and exonization.

If there existed a systematic way to study noisy splicing, it would be relatively straightforward to interpret if a subset of SNPs was acting to contribute to variation in splicing by reducing the production of functional transcripts and increasing the production of non-functional transcripts. Our hypothesis is that SNPs that increase disease risk through splicing often do so by reducing splicing accuracy, resulting in the production of a number of non-functional mRNA transcripts. Therefore, in this study, we developed a method to characterize noisy splicing or splicing fidelity, which will provide the field with a better understanding of the molecular mechanisms underlying sQTL activity and help in contextualizing sQTL effect sizes in the context of disease.

4.3 Results

4.3.1 Defining noisy splicing events

We worked with RNA-seq data from the panel of 364 Central European (CEU) lymphoblastoid cell lines (LCLs) from the GEUVADIS Consortium [34, 94] to identify noisy splicing events. In order to do this, we adapted a tool our lab routinely uses to study differential splicing events and map sQTL, LeafCutter [103]. In brief, LeafCutter identifies and quantifies annotated and novel splicing events by focusing on alternatively excised introns, and it does not rely on existing annotations. The intuition behind this "intron-centric" method is that mRNA splicing occurs through the step-wise removal of introns from nascent pre-mRNA, resulting in ligated exon-exon junctions in the mature mRNA. LeafCutter uses the junction reads that are captured from RNA-seq, which are representative of intron splicing or intron excision events, to identify all possible junctions. LeafCutter then groups together all overlapping intron excision events into clusters and assigns an intron excision ratio to every single intron excision event within a cluster (defined as the number of reads supporting that splicing event versus all splicing events in the same cluster). One can then perform differential splicing analysis to compare intron excision events between two different conditions or map sQTL by associating individual intron excision ratios with SNPs within a user-defined distance.

Two main approaches have been used to distinguish noisy splicing from functional splicing. First, because the RNA splicing process is associated with a small, but nonzero, error rate, one way to distinguish noisy splicing events from functional ones is on the basis rare usage (one caveat of LeafCutter is that it discards intron excision events that are observed $< 5\%$ of the time as compared to all other splicing events in the same cluster. Because noisy splicing events are expected to be observed at low frequencies, we did not filter such low-frequency intron excision events out of intron clusters). Second, because functional splic-

ing events are generally conserved across species, another way to distinguish noisy splicing events from functional ones is on the basis of low conservation across species.

We defined an intron or splicing junction to be noisy using the following three criteria: (i) the use of the intron relative to introns that share a splice site must be less than 0.1 in the analyzed sample, (ii) if the intron is detected in GTEx consortium [33] samples then its relative usage in GTEx samples from 54 tissues must be less than 0.1 in at least 95% of GTEx samples (the 5th percentile of usage must be less than 0.1), and (iii) the intron must not be annotated in more than two isoforms as annotated in the GENCODE V37 database. These parameters were chosen to reduce error rates in classifying a functional intron as noisy, and also to limit the number of noisy introns classified as functional.

We reasoned that noisy splicing events result in split RNA-seq reads demarcating introns that are rarely used and often not annotated. Thus, to establish a list of putative mis-spliced introns, we first searched for introns that are rarely used across samples. To do this, we clustered introns from samples from 54 tissues from the GTEx consortium and then estimated a relative usage of every intron based on the number of split reads supporting the intron relative to the number of split reads supporting other introns within the same cluster. As such, an intron is rarely used if it has a very low relative usage across all samples across all tissues (i.e. noisy introns are those with a 95th percentile usage less than 0.1). As expected, unannotated splice sites demarcating these rarely used introns are also depleted in evolutionarily conserved introns, consistent with being products of noisy splicing. Although our cutoff might be overly conservative, the number of introns classified to be mis-spliced exceeds that of annotated introns and likely represents true mis-spliced introns with no function.

We aimed to identify features that are predictive of an intron's level of splicing error. For example, we confirmed that longer introns exhibit higher levels of splicing error (Figure 4.1 A). In addition, highly expressed genes exhibit less error-prone splicing (Figure 4.1 B). These

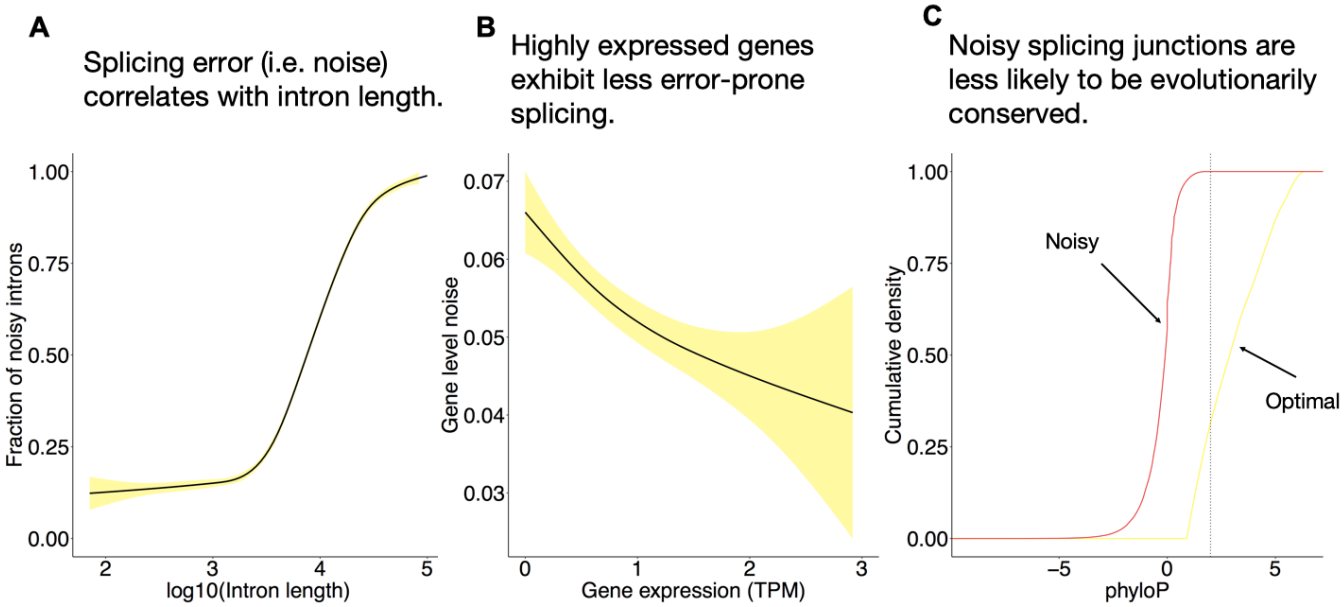


Figure 4.1: **Features of noisy splicing events.** (A) Splicing error, or noise, correlates with intron length. Introns were placed in100 bins based on length. In each bin, we calculated the mean fraction of sequencing reads from either splice site to an unconserved splice site and plotted against mean intron length. (B) Highly expressed genes exhibit less error-prone splicing. We assigned splicing junctions to genes and calculated transcripts per million (TPM) as (junction reads in gene) / (total junction reads) x 1 million.(C) Noisy splicing junctions are less likely to be evolutionarily conserved. We plotted the phyloP between the 3' and 5' splice sites. The gray vertical line is a phyloP cutoff of 1.75.

results are consistent with what a previous study of noisy splicing identified [137]. We also hypothesized that unannotated splice sites and their associated splicing junctions would be less likely to be evolutionarily conserved. In order to assess this, we compared the sequence conservation across placental mammals, using the phyloP score [140], between the noisy splicing junction, as defined previously, and functional or optimal splicing junctions. Indeed, we confirmed that noisy splicing junctions are less likely to be evolutionarily conserved (Figure 4.1 C).

4.3.2 The impact of inter-individual variation on splicing fidelity

To study inter-individual variation in alternative slicing and noisy splicing, we defined a noisy splicing event as the aggregate of all noisy introns within a cluster of introns. Specifically, we added the usage ratios per individual of all noisy introns in a cluster, quantile normalized, and used this as the phenotype in mapping splicing fidelity QTL (sfQTL). Controlling the false-discovery rate at 10%, we tested 82,843 noisy splicing events across 364 CEU LCLs and identified 3,269 sfQTL (within a 100kb window). Additionally, in the same cohort of individuals, we identified 11,138 sQTL (FDR$<$ 10%).

We hypothesized that sQTL could be explained by SNPs that affect canonical splice sites either by directly altering their strength or by affecting RNA binding protein or splicing enhancer motifs that regulate splice site choice. SNPs that disrupt canonical splice sites may increase noisy splicing. In other words, the expectation is that the production of non-functional, noisy transcripts may be a mechanism underlying a large fraction of sQTL, which is easily interpretable. An sfQTL with a positive effect size (i.e. a SNP that increases noisy splicing as a function of the number of minor alleles that an individual bears) that is also an sQTL with a negative effect size (i.e. a SNP that decreases how frequently a particular intron gets excised) can be interpreted mechanistically as a SNP that disrupts a canonical splice site, as described previously. Moreover, some SNPs that affect splicing through the

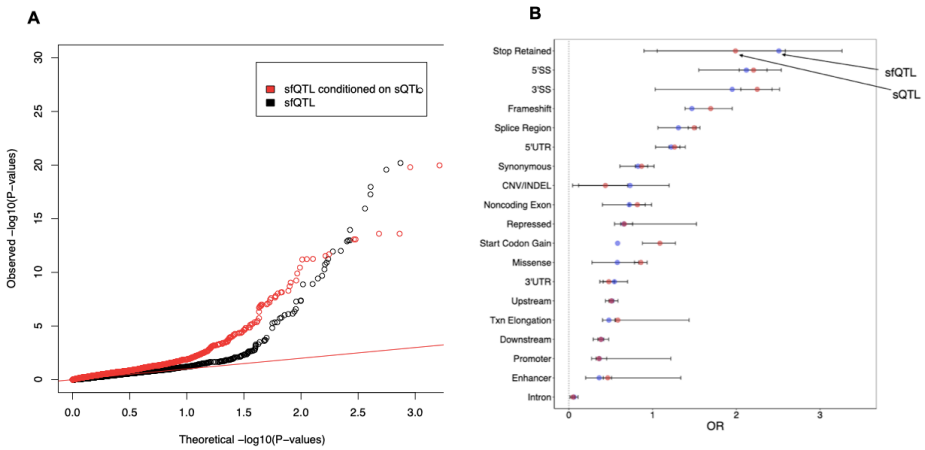


Figure 4.2: **Properties of sfQTL.** (A) QTLs for splicing fidelity are more likely to be sQTLs than matched SNPs within intron clusters. (B) Enrichment of QTL in genic and functional element annotations. Blue represents sfQTL. Red represents sQTL. Variant annotations were derived from SnpEff [30]. OR = odds ratio.

production of non-functional transcripts may also contribute to gene expression differences. For example, a SNP that increases the amount of noisy splicing (a sfQTL with a positive effect size) may enable the production of non-functional transcripts that are degraded by nonsense mediated decay (NMD), leading to decrease in gene expression (i.e. an eQTL with a negative effect size).

In order to begin to elucidate the mechanisms by which sfQTL might act, we used SnpEff [30] a variant annotation and effect prediction tool. In brief, Snpeff compares SNPs against known variants, such as those in dbSNP [156], and leverages genome annotations to predict information about SNPs, such as if they are over-represented in splice sites, alter amino acid sequences, and so forth. As expected, sfQTL were enriched in both 5' and 3' splice sites (Figure 4.2, OR = 2.1, 1.9, respectively). However, there was no statistically significant difference between sfQTL and sQTL enrichment in these variant annotations (Figure 4.2).

4.3.3 Understanding the role of noisy splicing in disease

We next aimed to identify examples of SNPs that impact splicing fidelity. As an example, we focused on the gene *IFI44L*, which is a type 1 interferon-stimulated gene known to inhibit human hepatitis virus replication [150]. Previously, others have demonstrated that *IFI44L* splicing is influenced by rs1333973, which can result in the inclusion or exclusion of the second exon of *IFI44L* [146]. Here, we determined that individuals with the G allele at rs273261 ($G > A$), which is within the first intron of the gene, exhibit impaired splicing fidelity (Figure 4.3 B, $P < 6.43e-89$; $\beta = -1.41$). In comparison, individuals with the alternative A allele splice out this intron and exhibit intact splicing fidelity. As expected, rs273261 is also a sQTL, with more A alleles resulting in increased intron excision of this first intron (Figure 4.3 C, $P < 1.55e-91$; $\beta = 1.31$). Indeed, this is an example in which *IFI44L* has an sQTL that is difficult to interpret initially given that the two splicing events are negatively correlated. In this case, individuals with more G alleles promote skipping of the

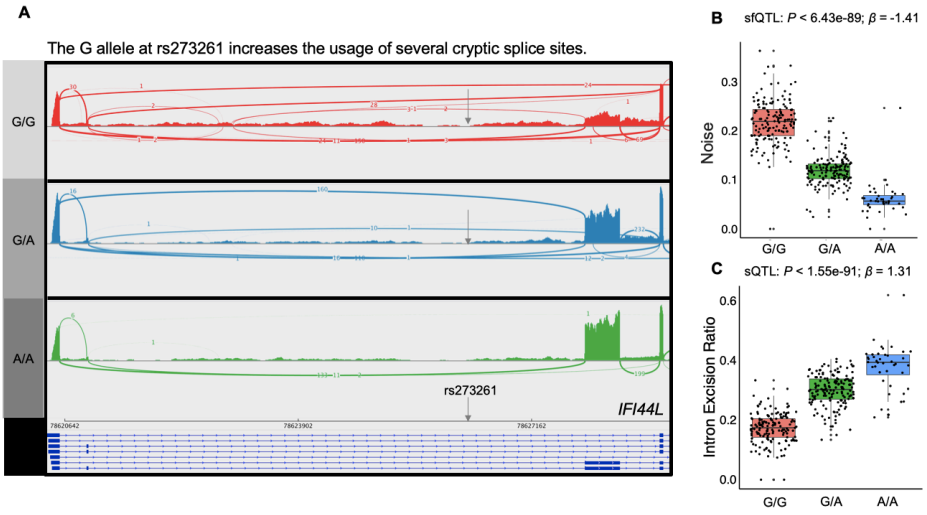


Figure 4.3: **Example of individuals with G alleles exhibiting a decrease in splicing fidelity but no decrease in gene expression.** (A) Shashimi plot of rs273261 (G > *A*) within *IFI44L*. (B) Boxplot of the representative sfQTL (sfQTL: P $< 6.43e-89$; $\beta = -1.41$). Individuals with the alternative A allele splice out the first intron in *IFI44L*. In contrast, individuals with the G allele exhibit impaired splicing fidelity and usage of cryptic splice sites within the first intron. (C) Boxplot of the associated sQTL (sQTL: P $< 1.55e-91$; $\beta = 1.31$).

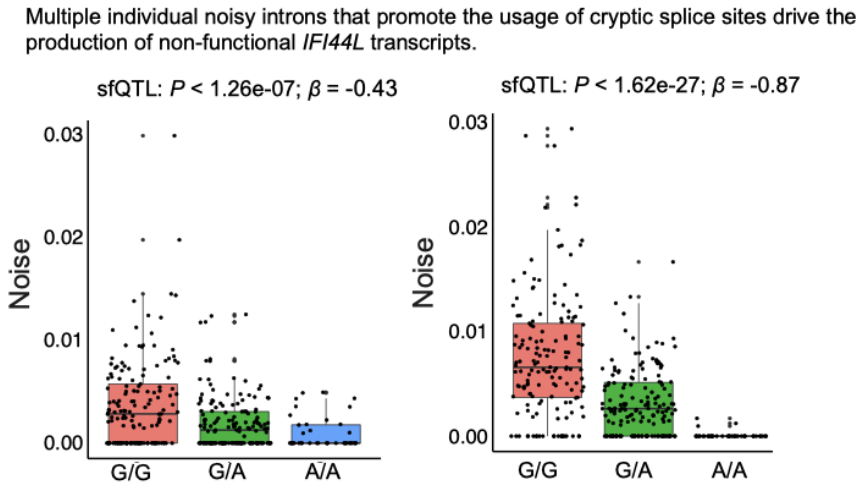


Figure 4.4: **rs273261 promotes the usage of multiple cryptic splice sites to impair splicing fidelity.** Multiple individual noisy introns that promote the usage of cryptic splice sites drive the production of non-functional IFI44L transcripts (two out of six introns). Boxplots of two representative sfQTL ($P < 6.43e-89$; $\beta = -1.41$ and $P < 1.55e-91$; $\beta = 1.31$).

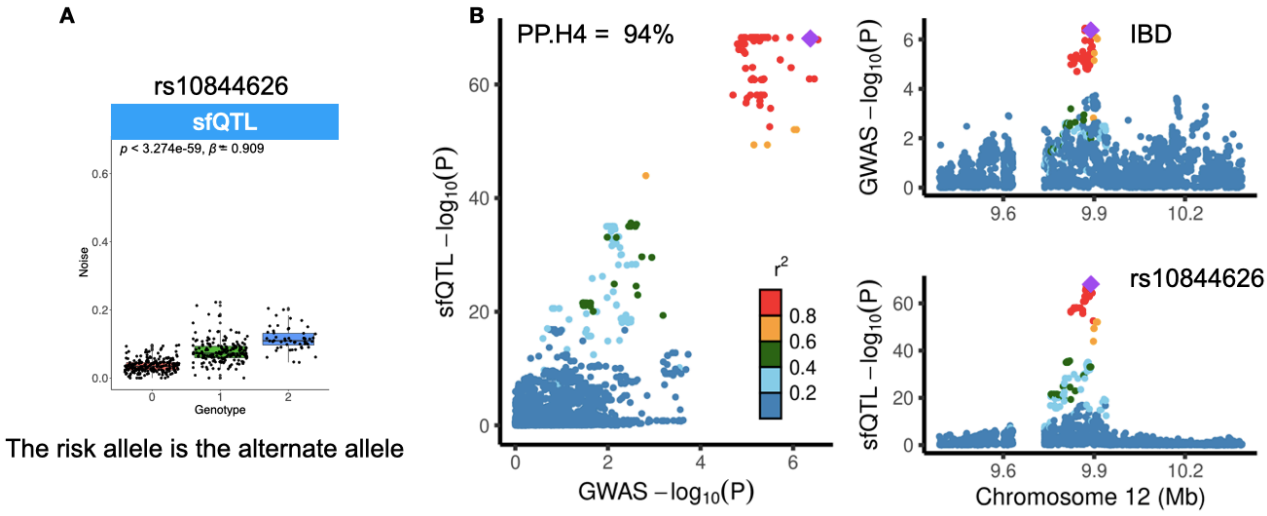


Figure 4.5: **rs10844626 is a shared casual variant underlying variation in splicing fidelity and IBD.** (A) Boxplot of a representative sfQTL, rs10844626 (T> A, C), in which the alternative allele promotes impaired splicing fidelity. (B) Posterior probability hypothesis 4 or PP.H4 ≤ 0.75 signifies that both the variant, in this case rs10844626, and the trait, in this case IBD, are associated and share a single causal variant.

second exon of *IFI44L*, but this appears to also be driven in part by the use of cryptic splice sites in the upstream exon, which would not have been identified without characterizing splicing fidelity.

In order to assess if an individual noisy intron or multiple noisy introns were driving splicing noise, we mapped sfQTL for all introns independently, or in a non-aggregated fashion. We discovered that two noisy introns out of six this cluster were driving the production of some non-functional *IFI44L* transcripts (Figure 4.4 C and Figure 4.4 D). Specifically, it appears as though rs273261 might be tagging a SNP in the first intron of *IFI44L* that either promotes usage of several cryptic splice sites in the middle of this *IFI411* intron, either directly or indirectly, or a SNP that promotes retention of the first intron (Figure 4.3 A).

In addition, we expected that in certain cases increasing the production of non-functional transcripts would also result in decreased gene expression. However, in this example, significant gene expression changes in *IFI44L* were not correlated with rs1333973 (nominal eQTL $P < 0.26$; $\beta = -0.07$).

To ascertain the role of noisy splicing in the context of disease, we compiled a set of 72 well-powered GWAS, including 14 for autoimmune diseases (11 unique disease types), 36 blood traits, and 22 other traits, as previously described [124], and used COLOC [59] to evaluate colocalization of GWAS hits and sfQTL signals. The idea behind approximate Bayes factor analysis is that the association between each trait with SNPs in a region may be summarized by a vector of 0s and, at most, a single 1. The single 1 is indicative of the causal variant or SNP, if we assume that a single causal variant underlies every trait. The posterior probability of each possible configuration can be calculated and so can the posterior probabilities that the traits share their configurations. Posterior probability hypothesis 4 (or PP.H4 ≤ 0.75) signifies that, in our case, a LCL-derived sfQTL and a trait share a single causal variant. Indeed, we identified an example of a shared causal variant, rs10844626, underlying variation in splicing fidelity and irritable bowel syndromes (IBD) (Figure 4.5).

4.3.4 Discussion and future directions

Li et al. previously showed that mRNA splicing is a primary mechanism that links genetic variation to disease. Several other studies have underscored this point by identifying sQTL in LD with GWAS hits for a number of diseases [104]. This suggests that SNPs affecting splicing have the potential to be causal variants underlying a substantial number of GWAS hits. However, one limitation is that the direction of effect for sQTL is difficult to interpret because alternatively spliced introns and exons are often correlated. Therefore, it is crucial to understand the mechanisms by which sQTL may disrupt function. We hypothesized that SNPs might increase disease risk by reducing splicing accuracy, resulting in the production of non-functional mRNA transcripts (i.e. noisy splicing). To test this hypothesis, we adapted LeafCutter to detect and map noisy splicing events and sfQTL. With this, we gained some intuition around the disruption of splicing accuracy as a potential mechanism underlying sQTL activity. In the future, it will be imperative to understand how sQTL and sfQTL act to contribute to disease phenotypes, which can be assessed through a variety of additional analyses.

We reported a systematic genetic analysis of splicing fidelity. Our analysis revealed insights into noisy splicing, highlighting that noisy splicing correlates with features such as intron length, gene expression, and sequence conservation. In addition, our analysis supported a role of splicing fidelity as a link between genetic variation and phenotypic variation, as evidenced by our integrated analysis of sfQTL with GWAS hits. We assessed enrichment of sfQTL in functional annotations to begin to characterize how sfQTL may act impact splice sites or RBPs. In the future, for each SNP within the sequence of an annotated splice site, we could compute a score based on how many differences there are between this modified splice site and the annotated splice site, such that a higher score reflects greater weakening of splice site strength. In addition, we could quantify enrichment of sfQTL within predicted RBP motifs from mCross for 112 RBPs from ENCODE eCLIP data (from K526 and HepG2

cell lines) [52] and use the DeepBind model [7] to predict the effect of these SNPs on RNA binding protein-RNA binding. DeepBind takes RNA sequences as input and outputs Deep-Bind scores, which quantify the binding specificity of different RBPs for the input sequences. DeepBind scores can be used to generate mutation maps, which visually display the impact of genetic variants on RBP-RNA binding. This would deepen our understanding of how sfQTL mechanisms of action.

To assess what fraction of sQTL can be explained by sfQTL and what fraction of SNPs that are sQTL and sfQTL are also eQTL, we could formally tested the sharing of effects between the two molecular phenotypes separately using Storey's π 1 statistic. We would expect that a large fraction of sQTL to be explained SNPs that affect noisy splicing, and approximately 25% of SNPs would also be eQTL. First, we could take the top SNPs for all sQTL and marked the associations between those SNPs and noisy splicing events. We could then compare the minimum P-values of the marked associations to the minimum p-value from 1,000 permutations of sample labels to compute an empirical p-value, which would represent the association between each sQTL and at least one noisy splicing event (q-value). From this, we could use Storey's π 0 statistic to compute the proportion of null associations, which is the sharing of sQTL and sfQTL.

It is worth noting that while we believe looking at the overlap between sQTL, sfQTL, and eQTL will be informative in understanding the mechanism of action for some SNPs associated with variation in splicing, evaluating the relative contribution of various other molecular phenotypes and pathways to RNA splicing variation will be important to obtain a more comprehensive understanding of sQTL mods of action. Though beyond the scope of this study, the following could be done. For example, in order to understand the relative contribution of chromatin phenotypes on variation in RNA splicing, one could quantify chromatin activity from ENCODE data to identify peaks of activity for various histone marks (e.g. H3K4me3, H3K27ac, H3K27me3, H3K36me3) and transcription factor binding.

I can then test whether these chromatin-level phenotypes are associated with variation in splicing, as Li et al. has done previously [104]. It is worth nothing that the authors identified a sQTL, rs6269 (A> G) that is associated with variation in CTCF binding and other aspects of chromatin, such as DNAse I hypersensitivity and methylation. However, the mechanism of action that affects splicing remains unclear. It is possible that by increasing CTCF binding, which has been shown to slow down RNA polymerase II transcription rate, may result in inefficient recognition of a canonical 3'SS, resulting in the use of a more upstream, alternative 3'SS. However, this putative mechanism would need to be tested experimentally with slow and fast RNA polymerase II mutants. For example, in G/G individuals expressing fast RNA polymerase II mutants, one would expect to see rescued use of the canonical 3'SS (i.e. in other words, the effect of this SNP on splicing would disappear).

We could further test the hypothesis that SNPs that increase disease risk through splicing likely do so by reducing splicing accuracy, resulting in the production of a number of non-functional mRNA transcripts, by performing transcriptome-wide association studies (TWASs), which have been developed to leverage expression data by imputing gene expression across a large cohort of genotyped individuals to identify target genes associated with disease phenotypes of interest [111]. In brief, for a given gene, TWAS is a test of local genetic association between gene expression and GWAS risk. Under a direct model in which a SNP acts to alter gene expression that leads to disease risk, TWAS will identify genes whose genetically regulated expression is associated with disease risk. Because we are interested in the role of splicing in disease, we could perform splicing TWAS and splicing fidelity TWAS to identify introns and noisy splicing events whose genetically regulated splicing is associated with disease risk. In order to perform splicing TWAS and noisy splicing TWAS, we could leverage RNA-seq data from a number of disease-relevant tissues to impute genetically-regulated splicing and noisy splicing across a large cohort of genotyped individuals. To do this, we would need to assess if there is high sharing of genetically-regulated splicing events

and noisy splicing events across disease-relevant tissues.

4.4 Materials and methods

4.4.1 Noising splicing events

We aligned and processed RNA-seq dataset from 4 species (rhesus, chimp, rat, and mouse) and calculated relative intron usage after lifting over to the human genome [72]. We reasoned that functional introns should be used at higher levels as compared to noisy introns, which should not be used in other species. Thus, we used the quantified intron usage in the 4 non-human species as a way to evaluate our noisy intron classification under different choices of parameters. We evaluated a range of possible cutoff parameters for noisy intron classification by asking about the number and fraction of introns in each category that have measurable usage in the four aforementioned species. The optimal parameters are such that intron classified as noisy are not used in other species, and intron classified as functional are used highly in other species. As expected, a more inclusive (lower) PSI cutoff for which introns are classified as functional result in a smaller percentage of them being used in other species. This is simply due to misclassification of noisy introns as functional. A more inclusive PSI cutoff for functional introns also result in a smaller fraction of introns classified to be noisy that are used in other species because it reduces the likelihood that functional introns are classified as noisy. Thus, our choice of parameter should reflect a balance that maximizes the number of introns classified as noisy, while minimizing the chance that we classify a functional intron as noisy. Using a PSI cutoff from 0.01 to 0.1 at 1-, 5-, 10-, 20- percentiles resulted in similar classification performance. Using our chosen PSI cutoff of 0.1 at the 5th percentile resulted in 50X more introns classified as functional that are used (> 0.1 PSI) in other species than compared to introns classified as noisy. Thus, 0.1 cutoff at the 5-percentile achieves our goal to identify noisy introns with minimal false positive function introns.

4.4.2 eQTL and sQTL mapping

Standard, short-read RNA-seq data for 364 CEU LCLs was obtained from the GEUVADIS project (EBI ArrayExpress, under the accession E-GEUV-2). Reads were mapped to the hg38 human reference genome using STARv2.6 [46], and WASP was used to filter out allele-specific reads that map with a bias [173]. We followed the QTL mapping pipelines that we established in previous studies of similar data [153]. In brief, in the context of eQTL analysis, to quantify gene expression levels, we used Kallisto [48] and added the transcript per million (TPM) estimates of all GENCODE v37 isoforms to obtain a gene-level TPMs. The gene-level TPMs were then scaled and quantile-quantile normalized as described previously [103]. We identified potential covariates by running principal component analysis (PCA) (prcomp function in R) and regressed out the top ten PCs using a linear model. In the context of sQTL analysis, normalized intron excision ratios calculated by LeafCutter [103] were used as phenotypes for sQTL mapping. Similarly, identified potential covariates by running principal component analysis (PCA) and regressed out the top ten PCs using a linear model.

QTLtools [41], an updated version of FastQTL [128], was used to test for association between SNPs within a cis-region of +-100kb of the gene or intron cluster and intron ratios within cluster (1000 permutations) and the phenotypes of interest. We used the VCF file from GEUVADIS (445 samples, GRCH38.20170514, with variants filtered at minor allele frequency, MAF < 0.01). Beta approximated permutation p values were then multiple test corrected using the q-value Storey and Tibshirani FDR correction. We defined eQTLs and sQTL at FDR $< 10\%$.

4.4.3 sfQTL mapping

In order to map splicing fidelity QTL (sfQTL), we defined a noisy splicing event as the aggregate of all unannotated introns, or noisy introns, within a cluster. Specifically, we added the usage ratios, per individual, of all noisy introns in a cluster, quantile normalized, and will

used this as the phenotype in sfQTL mapping. We identified potential covariates by running principal component analysis (PCA) and regressed out the top 10 PCs using a linear model. We tested the association between the phenotype described (i.e. an aggregated, noisy intron excision ratio per cluster) and all SNPs within 100kb of the intron cluster using QTLtools [41]. We also repeated this analysis on individuals noisy introns (i.e. non-aggregated). We defined sfQTLs at FDR $< 10\%$.

4.4.4 Colocalization analysis of molecular QTLs and GWAS variants

Our colocalization analysis was performed using the Approximate Bayes Factor (ABF) test implemented in software COLOC [59].

Coloc computes five posterior probabilities (PP0, PP1, PP2, PP3 and PP4), each corresponding to a hypothesis: H0: no association with either trait; H1: association with trait 1, not with trait 2; H2: association with trait 2, not with trait 1; H3: association with trait 1 and trait 2, two independent SNPs; H4: association with trait 1 and trait 2, one shared SNP. We ran coloc incorporated in the FUSION pipeline with default parameters (using the R function Fusion.assoc test.R in FUSION software with - -coloc P flag) and used PP4 to assess evidence of colocalization. We visualized the colocalization of sfQTL QTL and GWAS associations using LocusCompareR package (https://github.com/boxiangliu/locuscomparer).

APPENDIX A

ADDITIONAL PUBLICATIONS

Identification and quantification of splicing quantitative trait loci

Shah A., Li Y.I.

Springer. In:Shi X. (eds) eQTL Analysis: Methods and Protocols, 51-62 (2020)

Abstract

Most complex traits, including diseases, have a large genetic component. Identifying the genetic variants and genes underlying phenotypic variation remains one of the most important objectives of current biomedical research. Unlike Mendelian or familial diseases, which are usually caused by mutations in the coding regions of individual genes, complex diseases are thought to result from the cumulative effects of a large number of variants, of which, the vast majority are noncoding. Therefore, to discern the genetic underpinnings of a complex trait, we must first understand the impact of noncoding variation, which presumably affects gene regulation. In this chapter, we outline the recent progress made and methods used to discover putative regulatory regions associated with complex traits. We will specifically focus on mapping splicing quantitative trait loci (sQTL) using Yoruba samples from GEUVADIS as a motivating example.

Characterizing the major structural variant alleles of the human genome

Audano P.A., Sulovari A.A,Graves-Lindsay T.A., Cantsilieris S., Sorensen M., Welch A.E., Dougherty M.L., Nelson B.J., **Shah A.**, Dutcher S.K., Warren W.C., Magrini V., McGrath S.D., Li Y.I., Wilson R.K., and Eichler E.E

Cell. 176(3), 663-675 (2019)

Abstract

In order to provide a comprehensive resource for human structural variants (SVs), we generated long-read sequence data and analyzed SVs for fifteen human genomes. We sequence resolved 99,604 insertions, deletions, and inversions including 2,238 (1.6 Mbp) that are shared among all discovery genomes with an additional 13,053 (6.9 Mbp) present in the majority, indicating minor alleles or errors in the reference. Genotyping in 440 additional genomes confirms the most common SVs in unique euchromatin are now sequence resolved. We report a ninefold SV bias toward the last 5 Mbp of human chromosomes with nearly 55% of all VNTRs (variable number of tandem repeats) mapping to this portion of the genome. We identify SVs affecting coding and noncoding regulatory loci improving annotation and interpretation of functional variation. These data provide the framework to construct a canonical human reference and a resource for developing advanced representations capable of capturing allelic diversity.

References

[1] A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121–1130, 2015.

[2] 1958 Birth Cohort Controls Jones Richard W. 18 McArdle Wendy L. 18 Ring Susan M. 18 Strachan David P. 19 Pembrey Marcus 18 20, Type 1 Diabetes Clayton David G. 2 Dunger David B. 2 41 Nutland Sarah 2 Stevens Helen E. 2 Walker Neil M. 2 Widmer Barry 2 41 Todd John A. 2, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[3] Joshua D Aaker, Benayahu Elbaz, Yuwen Wu, Timothy J Looney, Li Zhang, Bruce T Lahn, and Brian Popko. Transcriptional fingerprint of hypomyelination in zfp191 null and shiverer (mbp shi) mice. *ASN neuro*, 8(5):1759091416670749, 2016.

[4] Salah E Abdel-Ghany, Michael Hamilton, Jennifer L Jacobi, Peter Ngam, Nicholas Devitt, Faye Schilkey, Asa Ben-Hur, and Anireddy SN Reddy. A survey of the sorghum transcriptome using single-molecule long reads. *Nature communications*, 7:11706, 2016.

[5] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome biology*, 12(2):1–14, 2011.

[6] Claudio R Alarcón, Hani Goodarzi, Hyeseung Lee, Xuhang Liu, Saeed Tavazoie, and Sohail F Tavazoie. Hnrnpa2b1 is a mediator of m6a-dependent nuclear rna processing events. *Cell*, 162(6):1299–1308, 2015.

[7] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[8] Frederick W Alt, Alfred LM Bothwell, Michael Knapp, Edward Siden, Elizabeth Mather, Marian Koshland, and David Baltimore. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mrnas that differ at their 3? ends. *Cell*, 20(2):293–301, 1980.

[9] Seyed Yahya Anvar, Guy Allard, Elizabeth Tseng, Gloria M Sheynkman, Eleonora de Klerk, Martijn Vermaat, Raymund H Yin, Hans E Johansson, Yavuz Ariyurek, Johan T den Dunnen, et al. Full-length mrna sequencing uncovers a widespread coupling between transcription initiation and mrna processing. *Genome biology*, 19(1):1–18, 2018.

[10] Takeshi Ara, Fabrice Lopez, William Ritchie, Philippe Benech, and Daniel Gautheret. Conservation of alternative polyadenylation patterns in mammalian genes. *BMC genomics*, 7(1):189, 2006.

[11] Ashraful Arefeen, Juntao Liu, Xinshu Xiao, and Tao Jiang. Tapas: tool for alternative polyadenylation site analysis. *Bioinformatics*, 34(15):2521–2529, 2018.

[12] Ashraful Arefeen, Juntao Liu, Xinshu Xiao, and Tao Jiang. Tapas: tool for alternative polyadenylation site analysis. *Bioinformatics*, 34(15):2521–2529, 2018.

[13] E. Beaudoing, S. Freier, J. R. Wyatt, J. M. Claverie, and D. Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Res*, 10(7):1001–10, 2000.

[14] Jordana T Bell, Athma A Pai, Joseph K Pickrell, Daniel J Gaffney, Roger Pique-Regi, Jacob F Degner, Yoav Gilad, and Jonathan K Pritchard. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome biology*, 12(1):1–13, 2011.

[15] Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[16] Dwight E Bergles and William D Richardson. Oligodendrocyte development and plasticity. *Cold Spring Harbor perspectives in biology*, 8(2):a020453, 2016.

[17] Kalen Berry, Jiajia Wang, and Q Richard Lu. Epigenetic regulation of oligodendrocyte myelination in developmental disorders and neurodegenerative diseases. *F1000Research*, 9, 2020.

[18] Joseph A Bokar, Mary Eileen Rath-Shambaugh, Rachael Ludwiczak, Prema Narayan, and Fritz Rottman. Characterization and partial purification of mrna n6-adenosine methyltransferase from hela cell nuclei. internal mrna methylation requires a multisubunit complex. *Journal of Biological Chemistry*, 269(26):17697–17704, 1994.

[19] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[20] Christopher D Brown, Lara M Mangravite, and Barbara E Engelhardt. Integrative modeling of eqtls and cis-regulatory elements suggests mechanisms underlying cell type specificity of eqtls. *PLoS genetics*, 9(8):e1003649, 2013.

[21] Helena Bujalka, Matthias Koenning, Stacey Jackson, Victoria M Perreau, Bernard Pope, Curtis M Hay, Stanlislaw Mitew, Andrew F Hill, Q Richard Lu, Michael Wegner, et al. Myrf is a membrane-associated transcription factor that autoproteolytically cleaves to directly activate myelin genes. *PLoS biology*, 11(8):e1001625, 2013.

[22] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O?Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[23] Enrico Cannavò, Nils Koelling, Dermot Harnett, David Garfield, Francesco P Casale, Lucia Ciglar, Hilary E Gustafson, Rebecca R Viales, Raquel Marco-Ferreres, Jacob F Degner, et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature*, 541(7637):402–406, 2017.

[24] John H Carson, Kimberly Worboys, Kevin Ainger, and Elisa Barbarese. Translocation of myelin basic protein mrna in oligodendrocytes requires microtubules and kinesin. *Cell motility and the cytoskeleton*, 38(4):318–328, 1997.

[25] Patrizia Casaccia and Gabriel Corfas. Introduction to the special issue on myelin plasticity in the central nervous system. *Developmental neurobiology*, 78(2):65, 2018.

[26] Jos Manuel Prez Caadillas and Gabriele Varani. Recognition of gurich polyadenylation regulatory elements by human cstf64 protein. *The EMBO journal*, 22(11):2821–2830, 2003.

[27] Zongliang Chen, Baobao Wang, Xiaomei Dong, Han Liu, Longhui Ren, Jian Chen, Andrew Hauck, Weibin Song, and Jinsheng Lai. An ultra-high density bin-map for rapid qtl mapping for tassel and ear architecture in a large f2 maize population. *BMC genomics*, 15(1):1–10, 2014.

[28] Anthony M Chomyk, Christina Volsko, Ajai Tripathi, Sadie A Deckard, Bruce D Trapp, Robert J Fox, and Ranjan Dutta. Dna methylation in demyelinated multiple sclerosis hippocampus. *Scientific reports*, 7(1):8696, 2017.

[29] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. *Nature genetics*, 49(4):600–605, 2017.

[30] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *fly*, 6(2):80–92, 2012.

[31] Benjamin L Clayton, Aaron Huang, Rejani B Kunjamma, Ani Solanki, and Brian Popko. The integrated stress response in hypoxia-induced diffuse white matter injury. *Journal of Neuroscience*, 37(31):7465–7480, 2017.

[32] David R Colman, Gert Kreibich, Alan B Frey, and David D Sabatini. Synthesis and incorporation of myelin polypeptides into cns myelin. *The Journal of cell biology*, 95(2):598–608, 1982.

[33] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[34] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

[35] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[36] Ryan A Coots, Xiao-Min Liu, Yuanhui Mao, Leiming Dong, Jun Zhou, Ji Wan, Xingqian Zhang, and Shu-Bing Qian. m6a facilitates eif4f-independent mrna translation. *Molecular cell*, 68(3):504–514, 2017.

[37] Darren A Cusanovich, Bryan Pavlovic, Jonathan K Pritchard, and Yoav Gilad. The functional consequences of variation in transcription factor binding. *PLoS genetics*, 10(3):e1004226, 2014.

[38] Alan Dabney, John D Storey, and GR Warnes. qvalue: Q-value estimation for false discovery rate control. *R package version*, 1(0), 2010.

[39] Lama Darbelli, Gillian Vogel, Guillermina Almazan, and Stéphane Richard. Quaking regulates neurofascin 155 expression for myelin and axoglial junction maintenance. *Journal of Neuroscience*, 36(14):4106–4120, 2016.

[40] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, et al. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[41] Olivier Delaneau, Halit Ongen, Andrew A Brown, Alexandre Fort, Nikolaos I Panousis, and Emmanouil T Dermitzakis. A complete tool set for molecular qtl discovery and analysis. *Nature communications*, 8(1):15452, 2017.

[42] Adnan Derti, Philip Garrett-Engele, Kenzie D MacIsaac, Richard C Stevens, Shreedharan Sriram, Ronghua Chen, Carol A Rohl, Jason M Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome research*, 22(6):1173–1183, 2012.

[43] Ronald Desrosiers, Karen Friderici, and Fritz Rottman. Identification of methylated nucleosides in messenger rna from novikoff hepatoma cells. *Proceedings of the National Academy of Sciences*, 71(10):3971–3975, 1974.

[44] D. C. Di Giammartino, K. Nishida, and J. L. Manley. Mechanisms and consequences of alternative polyadenylation. *Mol Cell*, 43(6):853–66, 2011.

[45] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[46] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[47] Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, et al. Topology of the human and mouse m6a rna methylomes revealed by m6a-seq. *Nature*, 485(7397):201–206, 2012.

[48] Yuheng Du, Qianhui Huang, Cedric Arisdakessian, and Lana X Garmire. Evaluation of star and kallisto on single cell rna-seq data alignment. *G3: Genes, Genomes, Genetics*, 10(5):1775–1783, 2020.

[49] Benayahu Elbaz and Brian Popko. Molecular control of oligodendrocyte development. *Trends in neurosciences*, 42(4):263–277, 2019.

[50] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

[51] João Fadista, Petter Vikman, Emilia Ottosson Laakso, Inês Guerra Mollet, Jonathan Lou Esguerra, Jalal Taneera, Petter Storm, Peter Osmark, Claes Ladenvall, Rashmi B Prasad, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences*, 111(38):13924–13929, 2014.

[52] Huijuan Feng, Suying Bao, Mohammad Alinoor Rahman, Sebastien M Weyn-Vanhentenryck, Aziz Khan, Justin Wong, Ankeeta Shah, Elise D Flynn, Adrian R Krainer, and Chaolin Zhang. Modeling rna-binding protein specificity in vivo by precisely registering protein-rna crosslink sites. *Molecular cell*, 74(6):1189–1204, 2019.

[53] Xin Feng, Lei Li, Eric J Wagner, and Wei Li. Tc3a: the cancer 3? utr atlas. *Nucleic acids research*, 46(D1):D1027–D1030, 2018.

[54] Gianluca Figlia, Daniel Gerber, and Ueli Suter. Myelination and mtor. *Glia*, 66(4):693–707, 2018.

[55] Michaela Frye, Bryan T Harada, Mikaela Behm, and Chuan He. Rna modifications modulate gene expression during development. *Science*, 361(6409):1346–1349, 2018.

[56] Ye Fu, Dan Dominissini, Gideon Rechavi, and Chuan He. Gene expression regulation mediated through reversible m6a rna methylation. *Nature Reviews Genetics*, 15(5):293–306, 2014.

[57] Nicole L Garneau, Jeffrey Wilusz, and Carol J Wilusz. The highways and byways of mrna decay. *Nature reviews Molecular cell biology*, 8(2):113–126, 2007.

[58] Consortium Genomes Project, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[59] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383, 2014.

[60] Andreas J Gruber, Ralf Schmidt, Andreas R Gruber, Georges Martin, Souvik Ghosh, Manuel Belmadani, Walter Keller, and Mihaela Zavolan. A comprehensive analysis of 3 end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein c on cleavage and polyadenylation. *Genome research*, 26(8):1145–1159, 2016.

[61] Kevin CH Ha, Benjamin J Blencowe, and Quaid Morris. Qapa: a new method for the systematic analysis of alternative polyadenylation from rna-seq data. *Genome biology*, 19(1):1–18, 2018.

[62] Kevin CH Ha, Benjamin J Blencowe, and Quaid Morris. Qapa: a new method for the systematic analysis of alternative polyadenylation from rna-seq data. *Genome biology*, 19(1):45, 2018.

[63] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):20, 2021.

[64] Siew Ping Han, Karin S Kassahn, Adam Skarshewski, Mark A Ragan, Joseph A Rothnagel, and Ross Smith. Functional implications of the emergence of alternative splicing in hnrnp a/b transcripts. *Rna*, 16(9):1760–1768, 2010.

[65] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[66] Li He and Q Richard Lu. Coordinated control of oligodendrocyte development by extrinsic and intrinsic signaling cues. *Neuroscience bulletin*, 29:129–143, 2013.

[67] P Cody He and Chuan He. m6a rna methylation: from mechanisms to therapeutic potential. *The EMBO journal*, 40(3):e105977, 2021.

[68] Ye He, Jeff Dupree, Ju Wang, Juan Sandoval, Jiadong Li, Huifei Liu, Yang Shi, Klaus Armin Nave, and Patrizia Casaccia-Bonnefil. The transcription factor yin yang 1 is essential for oligodendrocyte progenitor differentiation. *Neuron*, 55(2):217–230, 2007.

[69] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

[70] Marylens Hernandez and Patrizia Casaccia. Interplay between transcriptional control and chromatin regulation in the oligodendrocyte lineage. *Glia*, 63(8):1357–1375, 2015.

[71] Christina J Herrmann, Ralf Schmidt, Alexander Kanitz, Panu Artimo, Andreas J Gruber, and Mihaela Zavolan. Polyasite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic acids research*, 48(D1):D174–D179, 2020.

[72] Angela S Hinrichs, Donna Karolchik, Robert Baertsch, Galt P Barber, Gill Bejerano, Hiram Clawson, Mark Diekhans, Terrence S Furey, Rachel A Harte, Fan Hsu, et al. The ucsc genome browser database: update 2006. *Nucleic acids research*, 34(suppl_1):D590–D598, 2006.

[73] Keith S Hoek, Grahame J Kidd, John H Carson, and Ross Smith. hnrnp a2 selectively binds the cytoplasmic transport sequence of myelin basic protein mrna. *Biochemistry*, 37(19):7021–7029, 1998.

[74] Stephanie A Hoffmann, Deniz Hos, Melanie Küspert, Richard A Lang, Robin Lovell-Badge, Michael Wegner, and Simone Reiprich. Stem cell factor sox2 and its close relative sox3 have differentiation functions in oligodendrocytes. *Development*, 141(1):39–50, 2014.

[75] OW Howell, A Palser, A Polito, S Melrose, B Zonta, Christoph Scheiermann, AJ Vora, PJ Brophy, and R Reynolds. Disruption of neurofascin localization reveals early changes preceding demyelination and remyelination in multiple sclerosis. *Brain*, 129(12):3173–3185, 2006.

[76] JUN Hu, Carol S Lutz, Jeffrey Wilusz, and BIN Tian. Bioinformatic identification of candidate cis-regulatory elements involved in human mrna polyadenylation. *Rna*, 11(10):1485–1493, 2005.

[77] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115–121, 2015.

[78] Jimmy L Huynh, Paras Garg, Tin Htwe Thin, Seungyeul Yoo, Ranjan Dutta, Bruce D Trapp, Vahram Haroutunian, Jun Zhu, Michael J Donovan, Andrew J Sharp, et al. Epigenome-wide differences in pathology-free regions of multiple sclerosis–affected brains. *Nature neuroscience*, 17(1):121–130, 2014.

[79] Allan Jacobson and Stuart W Peltz. Interrelationships of the pathways of mrna decay and translation in eukaryotic cells. *Annual review of biochemistry*, 65(1):693–739, 1996.

[80] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer?s disease risk. *Nature genetics*, 51(3):404–413, 2019.

[81] Guifang Jia, YE Fu, XU Zhao, Qing Dai, Guanqun Zheng, Ying Yang, Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, et al. N 6-methyladenosine in nuclear rna is a major substrate of the obesity-associated fto. *Nature chemical biology*, 7(12):885–887, 2011.

[82] Bong-Seok Jo and Sun Shim Choi. Introns: the functional benefits of introns in genomes. *Genomics & informatics*, 13(4):112–118, 2015.

[83] Rahul N Kanadia, Karen A Johnstone, Ami Mankodi, Codrin Lungu, Charles A Thornton, Douglas Esson, Adrian M Timmers, William W Hauswirth, and Maurice S Swanson. A muscleblind knockout model for myotonic dystrophy. *science*, 302(5652):1978–1980, 2003.

[84] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009, 2010.

[85] Shengdong Ke, Endalkachew A Alemu, Claudia Mertens, Emily Conn Gantman, John J Fak, Aldo Mele, Bhagwattie Haripal, Ilana Zucker-Scharff, Michael J Moore, Christopher Y Park, et al. A majority of m6a residues are in the last exons, allowing the potential for 3? utr regulation. *Genes & development*, 29(19):2037–2053, 2015.

[86] Hanna Kedzierska and Agnieszka Piekie?ko-Witkowska. Splicing factors of sr and hnrnp families as regulators of apoptosis in cancer. *Cancer letters*, 396:53–65, 2017.

[87] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010.

[88] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[89] MinHyeok Kim, Bo-Hyun You, and Jin-Wu Nam. Global estimation of the 3? untranslated region landscape using rna sequencing. *Methods*, 83:111–117, 2015.

[90] Elijah Koreman, Xiaowei Sun, and Q Richard Lu. Chromatin remodeling and epigenetic regulation of oligodendrocyte myelination and myelin repair. *Molecular and Cellular Neuroscience*, 87:18–26, 2018.

[91] Anton Kratz and Piero Carninci. The devil in the details of rna-seq. *Nature biotechnology*, 32(9):882–884, 2014.

[92] Anton Kratz and Piero Carninci. The devil in the details of rna-seq. *Nature biotechnology*, 32(9):882, 2014.

[93] Leonid Kruglyak and Eric S Lander. A nonparametric approach for mapping quantitative trait loci. *Genetics*, 139(3):1421–1428, 1995.

[94] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[95] Shih-Han Lee, Irtisha Singh, Sarah Tisdale, Omar Abdel-Wahab, Christina S Leslie, and Christine Mayr. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, 561(7721):127–131, 2018.

[96] Ang Li, Yu-Sheng Chen, Xiao-Li Ping, Xin Yang, Wen Xiao, Ying Yang, Hui-Ying Sun, Qin Zhu, Poonam Baidya, Xing Wang, et al. Cytoplasmic m6a reader ythdf3 promotes mrna translation. *Cell research*, 27(3):444–447, 2017.

[97] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[98] Huiliang Li and William D Richardson. Genetics meets epigenetics: Hdacs and wnt signaling in myelin development and regeneration. *Nature neuroscience*, 12(7):815–817, 2009.

[99] Lei Li, Yipeng Gao, Fanglue Peng, Eric J Wagner, and Wei Li. Genetic basis of alternative polyadenylation is an emerging molecular phenotype for human traits and diseases. *BioRxiv*, page 570176, 2019.

[100] Lei Li, Yipeng Gao, Fanglue Peng, Eric J Wagner, and Wei Li. Genetic basis of alternative polyadenylation is an emerging molecular phenotype for human traits and diseases. *Available at SSRN 3351825*, 2019.

[101] Weimin Li, Wencheng Li, Rakesh S Laishram, Mainul Hoque, Zhe Ji, Bin Tian, and Richard A Anderson. Distinct regulation of alternative polyadenylation and gene expression by nuclear poly (a) polymerases. *Nucleic acids research*, 45(15):8930–8942, 2017.

[102] Weimin Li, Wencheng Li, Rakesh S Laishram, Mainul Hoque, Zhe Ji, Bin Tian, and Richard A Anderson. Distinct regulation of alternative polyadenylation and gene expression by nuclear poly (a) polymerases. *Nucleic acids research*, 45(15):8930–8942, 2017.

[103] Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of rna splicing using leafcutter. *Nature genetics*, 50(1):151–158, 2018.

[104] Yang I Li, Bryce Van De Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

[105] Steve Lianoglou, Vidur Garg, Julie L Yang, Christina S Leslie, and Christine Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development*, 27(21):2380–2396, 2013.

[106] Steve Lianoglou, Vidur Garg, Julie L Yang, Christina S Leslie, and Christine Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development*, 27(21):2380–2396, 2013.

[107] Xiaoyan Lin, Jill W Miller, Ami Mankodi, Rahul N Kanadia, Yuan Yuan, Richard T Moxley, Maurice S Swanson, and Charles A Thornton. Failure of mbnl1-dependent post-natal splicing transitions in myotonic dystrophy. *Human molecular genetics*, 15(13):2087–2097, 2006.

[108] Xiaoyan Lin, Jill W Miller, Ami Mankodi, Rahul N Kanadia, Yuan Yuan, Richard T Moxley, Maurice S Swanson, and Charles A Thornton. Failure of mbnl1-dependent post-natal splicing transitions in myotonic dystrophy. *Human molecular genetics*, 15(13):2087–2097, 2006.

[109] Jianzhao Liu, Yanan Yue, Dali Han, Xiao Wang, Ye Fu, Liang Zhang, Guifang Jia, Miao Yu, Zhike Lu, Xin Deng, et al. A mettl3–mettl14 complex mediates mammalian nuclear rna n6-adenosine methylation. *Nature chemical biology*, 10(2):93–95, 2014.

[110] Nian Liu, Katherine I Zhou, Marc Parisien, Qing Dai, Luda Diatchenko, and Tao Pan. N6-methyladenosine alters rna structure to regulate binding of a low-complexity protein. *Nucleic acids research*, 45(10):6051–6063, 2017.

[111] Justin M Luningham, Junyu Chen, Shizhen Tang, Philip L De Jager, David A Bennett, Aron S Buchman, and Jingjing Yang. Bayesian genome-wide twas method to leverage both cis-and trans-eqtl information through summary statistics. *The American Journal of Human Genetics*, 107(4):714–726, 2020.

[112] Tom Maniatis. Mechanisms of alternative pre-mrna splicing. *Science*, 251(4989):33–34, 1991.

[113] Mireya Marin-Husstege, Michela Muggironi, Aixiao Liu, and Patricia Casaccia-Bonnefil. Histone deacetylase activity is necessary for oligodendrocyte lineage progression. *Journal of neuroscience*, 22(23):10333–10345, 2002.

[114] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[115] Christine Mayr. Evolution and biological roles of alternative 3? utrs. *Trends in cell biology*, 26(3):227–237, 2016.

[116] Christine Mayr. Evolution and biological roles of alternative 3 utrs. *Trends in cell biology*, 26(3):227–237, 2016.

[117] Christine Mayr and David P Bartel. Widespread shortening of 3? utrs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, 2009.

[118] Graham McVicker, Bryce Van De Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749, 2013.

[119] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[120] Stanislaw Mitew, Curtis M Hay, Haley Peckham, Junhua Xiao, Matthias Koenning, and Ben Emery. Mechanisms regulating the development of oligodendrocytes and central nervous system myelin. *Neuroscience*, 276:29–47, 2014.

[121] Mithun Mitra, Elizabeth L Johnson, and Hilary A Coller. Alternative polyadenylation can regulate post-translational membrane localization. *Trends in cell & molecular biology*, 10:37, 2015.

[122] Briana E Mittleman, Sebastian Pott, Shane Warland, Tony Zeng, Zepeng Mu, Mayher Kaur, Yoav Gilad, and Yang Li. Alternative polyadenylation mediates genetic regulation of gene expression. *Elife*, 9:e57492, 2020.

[123] Sarah Moyon and Patrizia Casaccia. Dna methylation in oligodendroglial cells during developmental myelination and in disease. *Neurogenesis*, 4(1):e1270381, 2017.

[124] Zepeng Mu, Wei Wei, Benjamin Fair, Jinlin Miao, Ping Zhu, and Yang I Li. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome biology*, 22(1):122, 2021.

[125] Christina Müller, Nina M Bauer, Isabelle Schäfer, and Robin White. Making myelin basic protein-from mrna transport to localized translation. *Frontiers in cellular neuroscience*, 7:169, 2013.

[126] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[127] Timothy W Nilsen and Brenton R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[128] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2016.

[129] Franco Pagani and Francisco E Baralle. Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews Genetics*, 5(5):389–396, 2004.

[130] Christos Pantelis, George N Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O Perkins, Olli Pietiläinen, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[131] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

[132] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462–464, 2014.

[133] MARTHA L Peterson, ER Gimmi, and ROBERT P Perry. The developmentally regulated shift from membrane to secreted mu mrna production is accompanied by an increase in cleavage-polyadenylation efficiency but no measurable change in splicing efficiency. *Molecular and cellular biology*, 11(4):2324–2327, 1991.

[134] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[135] Simone Picelli, Omid R Faridani, sa K Bjrklund, Gsta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

[136] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[137] Joseph K Pickrell, Athma A Pai, Yoav Gilad, and Jonathan K Pritchard. Noisy splicing drives mrna isoform diversity in human cells. *PLoS genetics*, 6(12):e1001236, 2010.

[138] Anilkumar M Pillai, Courtney Thaxton, Alaine L Pribisko, Jr-Gang Cheng, Jeffrey L Dupree, and Manzoor A Bhat. Spatiotemporal ablation of myelinating glia-specific neurofascin (nfascnf155) in mice reveals gradual loss of paranodal axoglial junctions and concomitant disorganization of axonal domains. *Journal of neuroscience research*, 87(8):1773–1793, 2009.

[139] Xiao-Li Ping, Bao-Fa Sun, LU Wang, Wen Xiao, Xin Yang, Wen-Jia Wang, Samir Adhikari, Yue Shi, Ying Lv, Yu-Sheng Chen, et al. Mammalian wtap is a regulatory subunit of the rna n6-methyladenosine methyltransferase. *Cell research*, 24(2):177–189, 2014.

[140] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[141] Nick J Proudfoot. Ending the message: poly (a) signals then and now. *Genes & development*, 25(17):1770–1782, 2011.

[142] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[143] Towfique Raj, Yang I Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying-Chih Wang, Bernard Ng, Ishaan Gupta, Vahram Haroutunian, et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in alzheimer?s disease susceptibility. *Nature genetics*, 50(11):1584–1592, 2018.

[144] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[145] Emanuel Rosonina and James L Manley. From transcription to mrna: Paf provides a new path. *Molecular cell*, 20(2):167–168, 2005.

[146] Maxime Rotival, Hélène Quach, and Lluis Quintana-Murci. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nature communications*, 10(1):1671, 2019.

[147] Ian A Roundtree, Molly E Evans, Tao Pan, and Chuan He. Dynamic rna modifications in gene expression regulation. *Cell*, 169(7):1187–1200, 2017.

[148] Ian A Roundtree, Guan-Zheng Luo, Zijie Zhang, Xiao Wang, Tao Zhou, Yiquang Cui, Jiahao Sha, Xingxu Huang, Laura Guerrero, Phil Xie, et al. Ythdc1 mediates nuclear export of n6-methyladenosine methylated mrnas. *elife*, 6:e31311, 2017.

[149] Ute Schmidt, Eugenia Basyuk, Marie-Cécile Robert, Minoru Yoshida, Jean-Philippe Villemin, Didier Auboeuf, Stuart Aitken, and Edouard Bertrand. Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. *Journal of Cell Biology*, 193(5):819–829, 2011.

[150] John W Schoggins, Sam J Wilson, Maryline Panis, Mary Y Murphy, Christopher T Jones, Paul Bieniasz, and Charles M Rice. A diverse range of gene products are effectors of the type i interferon antiviral response. *Nature*, 472(7344):481–485, 2011.

[151] Marina M Scotti and Maurice S Swanson. Rna mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32, 2016.

[152] Daniel R Semlow and Jonathan P Staley. Staying on message: ensuring fidelity in pre-mrna splicing. *Trends in biochemical sciences*, 37(7):263–273, 2012.

[153] Ankeeta Shah and Yang I Li. Identification and quantification of splicing quantitative trait loci. *eQTL Analysis: Methods and Protocols*, pages 51–62, 2020.

[154] Ankeeta Shah, Briana E Mittleman, Yoav Gilad, and Yang I Li. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome biology*, 22(1):1–21, 2021.

[155] Diane L Sherman, Steven Tait, Shona Melrose, Richard Johnson, Barbara Zonta, Wendy B Macklin, Stephen Meek, Andrew JH Smith, David F Cottrell, Peter J Brophy, et al. Neurofascins are required to establish axonal domains for saltatory conduction. *Neuron*, 48(5):737–742, 2005.

[156] Stephen T Sherry, Minghong Ward, and Karl Sirotkin. dbsnp?database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research*, 9(8):677–679, 1999.

[157] Hailing Shi, Xiao Wang, Zhike Lu, Boxuan S Zhao, Honghui Ma, Phillip J Hsu, Chang Liu, and Chuan He. Ythdf3 facilitates translation and decay of n6-methyladenosine-modified rna. *Cell research*, 27(3):315–328, 2017.

[158] Irtisha Singh, Shih-Han Lee, Adam S Sperling, Mehmet K Samur, Yu-Tzu Tai, Mariateresa Fulciniti, Nikhil C Munshi, Christine Mayr, and Christina S Leslie. Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nature communications*, 9(1):1–16, 2018.

[159] Duncan J Smith, Charles C Query, and Maria M Konarska. ?nought may endure but mutability?: spliceosome dynamics and the regulation of splicing. *Molecular cell*, 30(6):657–666, 2008.

[160] Satoko Suzuki, Noriko Ayukawa, Chisa Okada, Masami Tanaka, Susumu Takekoshi, Yoko Iijima, and Takatoshi Iijima. Spatio-temporal and dynamic regulation of neurofascin alternative splicing in mouse cerebellar neurons. *Scientific Reports*, 7(1):11405, 2017.

[161] Steven Tait, Frank Gunn-Moore, J Martin Collinson, Jeffery Huang, Catherine Lubetzki, Liliana Pedraza, Diane L Sherman, David R Colman, and Peter J Brophy. An oligodendrocyte cell adhesion molecule at the site of assembly of the paranodal axo-glial junction. *The Journal of cell biology*, 150(3):657–666, 2000.

[162] Yoshio Takagaki, Rebecca L Seipelt, Martha L Peterson, and James L Manley. The polyadenylation factor cstf-64 regulates alternative processing of igm heavy chain premrna during b cell differentiation. *Cell*, 87(5):941–952, 1996.

[163] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications*, 8(1):14519, 2017.

[164] Courtney Thaxton, Anilkumar M Pillai, Alaine L Pribisko, Marilyne Labasque, Jeffrey L Dupree, Catherine Faivre-Sarrailh, and Manzoor A Bhat. In vivo deletion of immunoglobulin domains 5 and 6 in neurofascin (nfasc) reveals domain-specific requirements in myelinated axons. *Journal of Neuroscience*, 30(14):4868–4876, 2010.

[165] Bin Tian and Joel H Graber. Signals for premrna cleavage and polyadenylation. *Wiley interdisciplinary reviews: RNA*, 3(3):385–396, 2012.

[166] Bin Tian, Jun Hu, Haibo Zhang, and Carol S Lutz. A large-scale analysis of mrna polyadenylation of human and mouse genes. *Nucleic acids research*, 33(1):201–212, 2005.

[167] Bin Tian, Jun Hu, Haibo Zhang, and Carol S Lutz. A large-scale analysis of mrna polyadenylation of human and mouse genes. *Nucleic acids research*, 33(1):201–212, 2005.

[168] Bin Tian and James L Manley. Alternative polyadenylation of mrna precursors. *Nature reviews Molecular cell biology*, 18(1):18–30, 2017.

[169] H. Tilgner, F. Grubert, D. Sharon, and M. P. Snyder. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci U S A*, 111(27):9869–74, 2014.

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

[171] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[172] Todd J Treangen and Steven L Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36, 2012.

[173] Bryce Van De Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11):1061–1063, 2015.

[174] Markus C Wahl, Cindy L Will, and Reinhard Lührmann. The spliceosome: design principles of a dynamic rnp machine. *cell*, 136(4):701–718, 2009.

[175] Fay Wang, John Flanagan, Nan Su, Li-Chong Wang, Son Bui, Allissa Nielson, Xingyong Wu, Hong-Thuy Vo, Xiao-Jun Ma, and Yuling Luo. Rnascope: a novel in situ rna analysis platform for formalin-fixed, paraffin-embedded tissues. *The Journal of molecular diagnostics*, 14(1):22–29, 2012.

[176] R. Wang, R. Nambiar, D. Zheng, and B. Tian. Polya_db 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res*, 46(D1):D315–D319, 2018.

[177] Xiao Wang, Zhike Lu, Adrian Gomez, Gary C Hon, Yanan Yue, Dali Han, Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, et al. N 6-methyladenosine-dependent regulation of messenger rna stability. *Nature*, 505(7481):117–120, 2014.

[178] Xiao Wang, Boxuan Simen Zhao, Ian A Roundtree, Zhike Lu, Dali Han, Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He. N6-methyladenosine modulates messenger rna translation efficiency. *Cell*, 161(6):1388–1399, 2015.

[179] Sharon W Way, Joseph R Podojil, Benjamin L Clayton, Anita Zaremba, Tassie L Collins, Rejani B Kunjamma, Andrew P Robinson, Pedro Brugarolas, Robert H Miller, Stephen D Miller, et al. Pharmaceutical integrated stress response enhancement protects oligodendrocytes and provides a potential multiple sclerosis therapeutic. *Nature communications*, 6(1):6532, 2015.

[180] Jiangbo Wei, Fange Liu, Zhike Lu, Qili Fei, Yuxi Ai, P Cody He, Hailing Shi, Xiaolong Cui, Rui Su, Arne Klungland, et al. Differential m6a, m6am, and m1a demethylation mediated by fto in the cell nucleus and cytoplasm. *Molecular cell*, 71(6):973–985, 2018.

[181] Qinjie Weng, Ying Chen, Haibo Wang, Xiaomei Xu, Bo Yang, Qiaojun He, Weinian Shou, Yan Chen, Yujiro Higashi, Veronique van den Berghe, et al. Dual-mode modulation of smad signaling by smad-interacting protein sip1 is required for myelination in the central nervous system. *Neuron*, 73(4):713–728, 2012.

[182] Yi-Lan Weng, Xu Wang, Ran An, Jessica Cassin, Caroline Vissers, Yuanyuan Liu, Yajing Liu, Tianlei Xu, Xinyuan Wang, Samuel Zheng Hao Wong, et al. Epitranscriptomic m6a regulation of axon regeneration in the adult mammalian nervous system. *Neuron*, 97(2):313–325, 2018.

[183] Marvin Wickens, Philip Anderson, and Richard J Jackson. Life and death in the cytoplasm: messages from the 3 end. *Current opinion in genetics & development*, 7(2):220–232, 1997.

[184] Jocelyn Widagdo and Victor Anggono. The m6a-epitranscriptomic signature in neurobiology: from neurodevelopment to brain plasticity. *Journal of neurochemistry*, 147(2):137–152, 2018.

[185] Zheng Xia, Lawrence A Donehower, Thomas A Cooper, Joel R Neilson, David A Wheeler, Eric J Wagner, and Wei Li. Dynamic analyses of alternative polyadenylation from rna-seq reveal a 3-utr landscape across seven tumour types. *Nature communications*, 5:5274, 2014.

[186] Lin Xiao, David Ohayon, Ian A McKenzie, Alexander Sinclair-Wilson, Jordan L Wright, Alexander D Fudge, Ben Emery, Huiliang Li, and William D Richardson.

Rapid production of new oligodendrocytes is required in the earliest stages of motor-skill learning. *Nature neuroscience*, 19(9):1210–1217, 2016.

[187] Chao Xu, Ke Liu, Hazem Ahmed, Peter Loppnau, Matthieu Schapira, and Jinrong Min. Structural basis for the discriminative recognition of n6-methyladenosine rna by the human yt521-b homology domain family of proteins. *Journal of Biological Chemistry*, 290(41):24902–24913, 2015.

[188] Huan Xu, Yulia Dzhashiashvili, Ankeeta Shah, Rejani B Kunjamma, Yi-lan Weng, Benayahu Elbaz, Qili Fei, Joshua S Jones, Yang I Li, Xiaoxi Zhuang, et al. m6a mrna methylation is essential for oligodendrocyte maturation and cns myelination. *Neuron*, 105(2):293–309, 2020.

[189] Angli Xue, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E Kemper, Zhili Zheng, Loic Yengo, Luke R Lloyd-Jones, Julia Sidorenko, Yeda Wu, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications*, 9(1):1–14, 2018.

[190] Congting Ye, Yuqi Long, Guoli Ji, Qingshun Quinn Li, and Xiaohui Wu. Apatrap: identification and quantification of alternative polyadenylation sites from rna-seq data. *Bioinformatics*, 34(11):1841–1849, 2018.

[191] Ki-Jun Yoon, Francisca Rojas Ringeling, Caroline Vissers, Fadi Jacob, Michael Pokrass, Dennisse Jimenez-Cyrus, Yijing Su, Nam-Shik Kim, Yunhua Zhu, Lily Zheng, et al. Temporal control of mammalian cortical neurogenesis by m6a methylation. *Cell*, 171(4):877–889, 2017.

[192] Yanan Yue, Jianzhao Liu, and Chuan He. Rna n6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes & development*, 29(13):1343–1355, 2015.

[193] Sara Zaccara, Ryan J Ries, and Samie R Jaffrey. Reading, writing and erasing mrna methylation. *Nature reviews Molecular cell biology*, 20(10):608–624, 2019.

[194] Chunxia Zhang, Yusheng Chen, Baofa Sun, Lu Wang, Ying Yang, Dongyuan Ma, Junhua Lv, Jian Heng, Yanyan Ding, Yuanyuan Xue, et al. m6a modulates haematopoietic stem and progenitor cell specification. *Nature*, 549(7671):273–276, 2017.

[195] Jian Zhang, Yen K Lieu, Abdullah M Ali, Alex Penson, Kathryn S Reggio, Raul Rabadan, Azra Raza, Siddhartha Mukherjee, and James L Manley. Disease-associated mutation in srsf2 misregulates splicing by altering rna-binding affinities. *Proceedings of the National Academy of Sciences*, 112(34):E4726–E4734, 2015.

[196] Ye Zhang, Kenian Chen, Steven A Sloan, Mariko L Bennett, Anja R Scholze, Sean O'Keeffe, Hemali P Phatnani, Paolo Guarnieri, Christine Caneda, Nadine Ruderisch, et al. An rna-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *Journal of Neuroscience*, 34(36):11929–11947, 2014.

[197] Zijie Zhang, Kaixuan Luo, Zhongyu Zou, Maguanyun Qiu, Jiakun Tian, Laura Sieh, Hailing Shi, Yuxin Zou, Gao Wang, Jean Morrison, et al. Genetic analyses support the contribution of mrna n 6-methyladenosine (m6a) modification to human disease heritability. *Nature genetics*, 52(9):939–949, 2020.

[198] Boxuan Simen Zhao and Chuan He. ?gamete on? for m6a: Ythdf2 exerts essential functions in female fertility. *Molecular cell*, 67(6):903–905, 2017.

[199] Xianghui Zhao, Jinxiang Dai, Yue Ma, Yajing Mi, Daxiang Cui, Gong Ju, Wendy B Macklin, and Weilin Jin. Dynamics of ten-eleven translocation hydroxylase family proteins and 5-hydroxymethylcytosine in oligodendrocyte differentiation. *Glia*, 62(6):914–926, 2014.

[200] Xianghui Zhao, Xuelian He, Xiaolei Han, Yang Yu, Feng Ye, Ying Chen, ThaoNguyen Hoang, Xiaomei Xu, Qing-Sheng Mi, Mei Xin, et al. Microrna-mediated control of oligodendrocyte differentiation. *Neuron*, 65(5):612–626, 2010.

[201] Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min Huang, Charles J Li, Cathrine B Vågbø, Yue Shi, Wen-Ling Wang, Shu-Hui Song, et al. Alkbh5 is a mammalian rna demethylase that impacts rna metabolism and mouse fertility. *Molecular cell*, 49(1):18–29, 2013.

[202] Jun Zhou, Ji Wan, Xiangwei Gao, Xingqian Zhang, Samie R Jaffrey, and Shu-Bing Qian. Dynamic m6a mrna methylation directs translational control of heat shock response. *Nature*, 526(7574):591–594, 2015.

[203] Jun Zhou, Ji Wan, Xin Erica Shu, Yuanhui Mao, Xiao-Min Liu, Xin Yuan, Xingqian Zhang, Martin E Hess, Jens C Brüning, and Shu-Bing Qian. N6-methyladenosine guides mrna alternative translation during integrated stress response. *Molecular cell*, 69(4):636–647, 2018.

[204] Qiao Zhou and David J Anderson. The bhlh transcription factors olig2 and olig1 couple neuronal and glial subtype specification. *Cell*, 109(1):61–73, 2002.

[205] Tingting Zhu, Ian A Roundtree, Ping Wang, Xiao Wang, Li Wang, Chang Sun, Yuan Tian, Jie Li, Chuan He, and Yanhui Xu. Crystal structure of the yth domain of ythdf2 reveals mechanism for recognition of n6-methyladenosine. *Cell research*, 24(12):1493–1496, 2014.

[206] Barbara Zonta, Steven Tait, Shona Melrose, Heather Anderson, Sheila Harroch, Jennifer Higginson, Diane L Sherman, and Peter J Brophy. Glial and neuronal isoforms of neurofascin have distinct roles in the assembly of nodes of ranvier in the central nervous system. *The Journal of cell biology*, 181(7):1169–1177, 2008.