THE UNIVERSITY OF CHICAGO

THE ROLE OF ALTERNATIVE POLYADENYLATION VARIATION IN GENE REGULATION DIFFERENCES WITHIN AND BETWEEN SPECIES

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, GENOMIC, AND SYSTEMS BIOLOGY

BY
BRIANA ERIN MITTLEMAN

CHICAGO, ILLINOIS
AUGUST 2020

"'Cause baby you're a firework

Come on show 'em what your worth

Make 'em go "Oh, oh, oh!"

As you shoot across the sky-y-y"

Katy Perry - Firework, 2010.

Table of Contents

List of Figures

List of Tables [1]

1. Note: Due to the large size of some tables, the tables have been provided in a supplementary file accompanying the dissertation. In such cases, the page number provided below directs the reader to a table's caption.

ACKNOWLEDGMENTS

The following dissertation would not have been possible without the immense amount of support that I have had throughout my time at the University of Chicago. It is difficult to put into words the gratitude that I have for all of the people that I have met here in Chicago.

First, I would like to thank my advisor, Dr. Yoav Gilad. Yoav has continued to be my advocate scientifically and personally. I am thankful both for how he has pushed me to become a better scientist while also providing me the intellectual freedom to pursue my goals. I hope that in the future I can continue to call Yoav both a mentor and a friend. I was also lucky to work with Dr. Yang Li as he started his lab at University of Chicago. Working with him and members of his lab has taught me about being a collaborator, mentor, and mentee. I am also forever grateful to have Dr. Sebastian Pott as an additional mentor in my lab. Seb is both a thoughtful scientist and a great mentor. On a day to day basis, I knew I could count on Seb for both scientific and person advice. I would like to thank my other committee members, Dr. Matthew Stephens and Dr. Jon Staley. They both provided an outside prospective and helped me to think about my work from a different angle.

From my first days at University of Chicago, the Gilad lab has been my home. Each and every member of the Gilad lab has made this experience worth it! I am so lucky to have started in the lab with 2 amazing women, Katie Rhodes and Reem Elorbany. While we cannot be any more different as people, our friendships are ones I will cherish forever. We will forever be the 'cohort' and remember to work 'deep in the hood'. Lauren Blake deserves a special thank you because she acted as my mentor from the day I interviewed until the day I defended. I learned more on our walks home than I did in many classes! I am so glad we will both be in California for the next few years. Thank you to John Blischak for teaching me the importance of reproducible science and ensuring that I learned and implemented best practices for data science as soon as I rotated in the Stephens lab. I also have to thank John for helping me get involved with Software Carpentry. Each Gilad lab member past and

present, has influenced my scientific and personal views. The other graduate students, Bryan Pavlovic, John Blischak, Lauren Blake, Ittai Eres, Katie Rhodes, Reem Elorbany, Anthony Hung, Deji Adegunsoye, Wenhe Lin and Erik McIntire. The post-docs Po-Yuan Tung, Joyce Hsiao, Michelle Ward, Genevieve Housman, Kenneth Barr, Ben Fair, Ben Umans. The lab managers and technicians, Kristen Patterson, Jonathan Burnett, Claudia Cuevas, and Emilie Briscoe. Also special thanks to our science writer Natalia Gonzales. I have also had to great opportunity to work with members of the Li lab including graduate student, Phoenix Mu, and the undergrads Tony Zeng, and Shane Warland.

I have also made some wonderful friends in the GGSB and HG program. Among others, Sahar Mozaffari, Linsin Smith, Sammy Keyport, Ryan Dohn, Michael Drazer, and Manny Vazquez. Other members of the research community, Ezra Amiri, Edgar Correa, Tomasz Slezak, and Daniel Downie. Thank you to Helen Robertson for her friendship and for providing useful feedback on chapters 1 and 5 of this dissertation. I need to give a special thank you to Haley Randolph, who also became my only roommate in Chicago. Haley works harder than anyone I know but still has time to be a true friend! I know she will solve big problems in Genomics and Immunology one day and I plan to be her biggest fan. Sue Levison has been way more than just a graduate program administrator during my time in Chicago. Sue has been my biggest fan and friend from the moment I got to Chicago. I value her advice and emotional support over the last 4 years. Scientific outreach and community engagement helped me stay grounded during my Ph.D. I would like to thank Shaz Rasul, Monica Luna, and the rest of the Neighborhood Schools Program team for backing my SMART Science program. Also, thank you to the students who helped make sure the program has been and will continue to be successful.

My success in this program would not have been possible without the communities outside of the University that have provided me a home in Chicago. First, Trapeze School New York in Chicago. Not only has trapeze given me amazing friends and a super unique hobby, it

gave me a reason to branch out of my community and take a break each week. I will never forget the Sunday's that I spent working at the Edgewater Starbuck's then flying in my IFW. Second, Orange Theory Fitness Hyde Park became more than a gym to me over the last few years. Through the OTF HP community I was able to rediscover my love for fitness. There are many stressful lab days that would have been a lot harder if I did not have my OTF workout to de-stress!

I made many lifetime friends as an undergraduate at Duke University. They have continued to be my biggest cheerleaders while I have been here in Chicago. First, thank you to my friends who lived in Chicago for at least a short time when I was here, Katie Heckman, Lauren Rosen, Adriana Dickerson, and Amanda Jones. I have to thank them for making the trek down to Hyde Park to come spend time with me sometimes! Also thank you to my friends who supported me from afar, Breanna Atkinson, Linda Zambrano, Sonia Lee. Also thank you the rest of my Duke Cheerleading class of 2016 family. Each of these people and so many more, mean so much to me and I am grateful that our friendships have stayed strong since graduation from Duke.

I would not be submitting this Ph.D. thesis without the continued support of Dr. Mohamed Noor. I will forever owe him for providing me the opportunity to start working in his lab. I am so lucky to have an undergraduate research advisor who will give me advice in hard times and be the first to congratulate me on achievements.

Finally, I need to thank my family. Even from across the country my mom- Lisa, dad- David, sister- Ariel, and brother-Bradley have continued to be my rocks. While they never understood what I was studying they still were happy to hear about the highs and lows of my research. I am also extremely lucky to have 4 supportive grandparents, Grandma Nancie, Papa Len, Grandma Linda, and Grandpa Richard. They each had the opportunity to visit me in Chicago during my graduate work. I hope to continue to make my family proud in the years to come.

ABSTRACT

Differences in gene regulation contribute to phenotypic differences within humans and also between humans and other primates. While co-transcriptional gene regulatory mechanisms such as alternative polyadenylation *(APA)* can help explain how variation in gene regulation manifests, such mechanisms remain understudied. In this thesis, I used a quantitative genomics approach and a comparative primate approach to understand the regulatory role of APA. In chapter 2, I measured polyadenylation site *(PAS)* usage genome wide in a population of 52 human lymphoblastoid cell lines. I identified genetic variation associated with APA *(apaQTLs)* and showed that genetic variation acts through PAS choice to impact mRNA expression, translation, and protein levels in complex, non-linear ways. In chapter 3, I measured APA conservation between human and chimpanzee. While APA is largely conserved, differences in PAS usage and isoform diversity contribute to differentially expressed and differentially translated genes. Together, these chapters, establish APA as a key co-transcriptional mechanism underlying the genetic regulation of gene and protein expression levels. As a step to further understand co-transcriptional regulatory mechanisms, in chapter 4, I describe an attempt to measure polymerase II elongation rate genome wide. In the final chapter, I outline a set of necessary future directions to extend my work on APA to more tissues and biological processes.

(Note: Supplementary tables are provided in a .zip file available online. Captions for the tables are provided within the dissertation.)

CHAPTER 1

INTRODUCTION

An overarching goal in human genetics is to understand the relationship between genotype and phenotype. An integral step toward this goal is learning how to interpret the functional consequences of genetic variants. In protein-coding regions, it is relatively easy to predict if a single nucleotide polymorphism *(SNP)* will affect protein sequence because the amino acid code is known. However, protein-coding genetic sequence makes up only $\sim 1\%$ of the genome. Although non-coding regions of the genome were categorized as "junk DNA" for decades, it is now clear that a large proportion of the non-coding genome is essential. Indeed, non-coding DNA has been established to play a large role in gene regulation - the chain of events controlling how and when cells express particular genes. Thus, understanding how non-coding genetic variants influence gene regulation is a vital step toward understanding the relationship between genotype and phenotype.

In this chapter, I will introduce the two approaches that I have taken in my work to study gene regulation. By providing examples, primarily from the Gilad lab, I will motivate why I have chosen to study gene regulation through quantitative genetics and comparative primate functional genomics. Following this discussion, I argue that co-transcriptional gene regulatory mechanisms remain understudied and warrant particular attention by both approaches. I conclude the chapter with an overview of the remaining chapters in this dissertation.

1.1 Quantitative genetic approaches to understand gene regulation

Quantitative trait loci *(QTL)* mapping is a tool used to identify genetic variation affecting gene regulation. For example, an expression QTL, or eQTL, is a genetic variant that is statistically associated with mRNA expression. Over the last decade, a large number of

eQTL mapping studies have identified genetic variants with regulatory potential in a wide range of human cell lines and tissues [89, 171, 171, 144, 61, 162]. In the most comprehensive project to date, the Genome-Tissue Expression *(GTEx)* Consortium generated eQTL maps for over 50 human tissues collected from over 800 healthy individuals [192, 2, 61]. Integration of these maps has revealed numerous eQTLs for the majority of protein coding genes.

While eQTLs have been largely successful in identifying SNPs with regulatory potential, this approach does not reveal the chain of events controlling gene expression levels. eQTLs represent a collection of SNPs mediating gene expression along the gene regulatory cascade through a wide range of mechanisms 1.1 [138]. The development of methods like DNAse-seq [170] ChIP-Seq [9], and ATAC-seq [24], which are used to identify chromatin states, made it possible to study the contribution of specific molecular mechanisms to variation in gene expression levels. For example, if we assume an eQTL acts by perturbing a particular mechanism (i.e. chromatin conformation), the SNP should also be associated with measures of that mechanism, and molecular QTL mapping further contributes to the understanding of gene regulation.

1.2 Comparative approaches to understand gene regulation

Comparative primate functional genomic studies have also contributed to our understanding of human gene regulation. Like phenotypic variation between individual human, phenotypic differences between human and non-human primates stem largely from differences in gene regulation [82, 30, 57, 196, 18, 77]. In turn, primate functional genomic studies complement human-specific quantitative genetic studies. Due to the challenges associated with obtaining primate tissues, comparative studies in primates typically rely on smaller sample sizes. However, these are sufficient to detect the large differences in regulatory mechanisms between species. Comparative genomic approaches also contextualize gene regulatory mechanisms in an evolutionary framework to provide additional insight into which processes or loci are

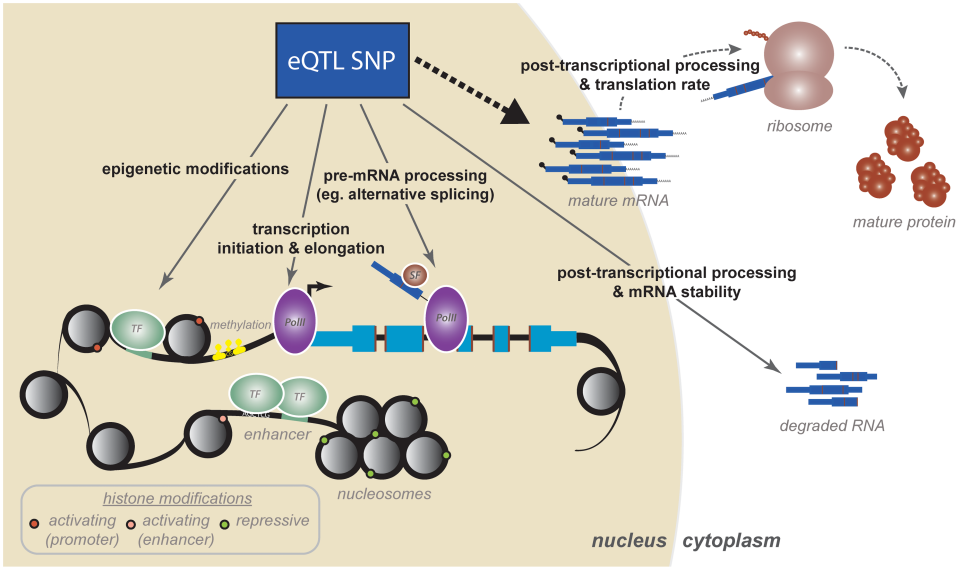


Figure 1.1: Figure 1 from Pai, Pritchard, and Gilad 2015 [138] Regulatory QTL studies have revealed how eQTLs can affect variation in mature mRNA expression levels as well as how variation in mRNA expression may then impact post transcriptional and translational mechanisms.

critical to genome and organismal function.

There are some regulatory processes that, while functionally important, are indistinguishable in human populations or are measured with technology not feasible to use at the scale necessary for quantitative genetic studies. Because genetic variation is larger between species than within human populations, it is possible to distinguish differences in regulatory features between species - even with small sample sizes. For example, before it became technically and financially feasible to measure chromosome contacts genome-wide at a human population level, Eres *et al.* demonstrated that differences in contact frequencies at individual loci between human and chimpanzee can explain previously identified differentially expressed genes [49]. Specifically, Eres *et al.* used Hi-C to characterize 3D chromatin contacts in human and chimpanzee induced pluripotent stem cells *iPSCs*. While a quantitative genetics study to detect genetic variation associated with significant differences in chromatin interactions between humans was not feasible, this work demonstrated that chromatin contact variation likely mediates differences in gene expression, and complemented previous work

that suggested direct contact between distal enhancer elements and promoters driving gene expression [31, 86, 182, 139, 109].

Comparative primate genome studies also provide evolutionary context for genetic mechanisms and genomic loci previously studied with alternative approaches. Under the assumption that conserved processes and loci are critical to function, understanding evolutionary context of gene regulatory mechanisms can help catalog importance of each mechanism [66]. For example, human quantitative genetics approaches revealed higher variation in mRNA transcript levels than protein levels, which likely result from mechanisms downstream of translation buffer protein levels [11]. However, the importance of post-translational gene regulation for maintaining protein level variation was unclear. Wang *et al*, measured translation in a set of primate LCLs with available mRNA expression and protein levels. Using these data, the authors demonstrated that post-translation protein buffering likely plays a large role in shaping primate protein level divergence [191]. Overall, characterizing gene regulatory mechanisms in both humans and non-human primates is critical to understanding the broader relationship between genotype and phenotype.

1.3 Open questions in how isoform diversity contributes to gene regulation

Both human population genomic and primate comparative genomic approaches have improved our ability to interpret noncoding variation, yet more work is necessary to understand the gene regulatory code. It has become clear that a large number of highly conserved genetic variants that correlate with variation in both mRNA expression and complex traits lie in cis-regulatory elements such as promoters and enhancers [110, 184, 207]. While considerable effort has been devoted to annotating promoters and enhancers across cell types and species [14, 176, 123, 119], a large proportion of variants associated with gene expression variation and/or complex traits remain unannotated and unexplained.

Indeed, the majority of previously explored gene regulatory mechanisms affect gene expression levels pre-transcriptionally through chromatin level variation at promoters and enhancers. However, co-transcriptional mechanisms, such as alternative mRNA splicing and alternative polyadenylation *(APA)*, contribute to both variation in mRNA and protein isoforms dependent on and independent of changes in mRNA expression levels. Understanding the co-transcriptional mechanisms that mediate isoform diversity is a key path forward to predicting phenotype from genotype. Thus, a greater understanding of how genetic variation and regulatory elements control alternative splicing and APA will fill a gap in predictive models of gene regulation from DNA sequences.

In recent years, a number of groups have started to identify and characterize the genetic variation associated with alternative splicing *(sQTLs)* [175, 95, 133, 103, 181]. In one such study, Li *et al.* mapped sQTLs in a population of LCLs that was previously used to characterize a number of other molecular QTLs. They reported that most sQTLs, including some previously implicated in GWAS, are not eQTLs [95]. This work suggests genetic variation can mediate traits independently from differences in mRNA expression levels.

Although comparative functional genomic studies of alternative splicing are still rare, a few studies have generated important insights. Blekhman *et al.* reported that genes with human-specific exon usage are enriched in genes involved in anatomical structure [17]. Additional work has demonstrated how alternative splicing may contribute specifically to differential expression and translation of genes between closely related primates [100, 6]. For example, Attig *et al.* reported that co-evolution of splice sites and U-rich elements in primates has led to an increase in expression of exons containing Alu transposable elements. Their experiments demonstrated inclusion of Alu exons causes down regulation of transcripts through non-sense mediated decay [6].

Like alternative splicing, cleavage and polyadenylation in introns leads to differential inclusion of coding sequences. mRNA transcripts resulting from intronic polyadenylation are

either subject to decay or are translated to distinct protein isoforms (reviewed in Tian and Manley [179]). Unlike alternatively spliced transcripts, APA in 3' untranslated regions (3' UTRs) produces transcripts with the same coding sequences thus only differing in 3' UTR length. Variation in downstream regulation of 3' UTR APA isoforms result from differential inclusion of miRNA binding sites and other motifs signaling RNA binding proteins. Thus, isoforms with distinct 3' UTRs are subject to alternative stability, localization, and translation (reviewed in Tian and Manley [179]).

While the molecular basis for cleavage and polyadenylation of mRNA transcripts is known, the precise mechanisms that discriminate between potential polyadenylation sites *(PAS)* in the same gene are not. Molecular biology studies have characterized the DNA motifs and the cascade of events leading to cleavage and polyadenylation of mRNA molecules [174, 36, 104]. In short, elongating RNA polymerase II recognizes a canonical signal motif *(AATAAA)* and recruits the necessary protein complexes to cleave the mRNA transcript and add adenosine bases to the 3' end. In turn, it is not surprising that studies have identified individual loci whereby signal site changes contribute to variation in APA and downstream regulation. However, over 80% of human genes that have multiple isoforms do not contain signal sites changes that could directly point to differential usage of individual isoforms.

Variation in APA has been linked to functional regulatory changes that contribute to differences in complex traits and disease risk. For example, a genetic variant in the signal site of a PAS in the *IRF5* gene has been causally linked to risk of systemic lupus erythematosus *(SLE)* [59]. Individuals with SLE have a genetic mutation that reduces the use of the most 5' proximal PAS in the 3' UTR of *IRF5*. Usage of the distal PAS produces a longer 3' UTR with an AU-rich destabilizing element *(ARE)*. As a result, the alternative isoform is quickly degraded, reducing gene expression of IRF5 in these patients. However, the role of APA in gene regulation and disease risk genome-wide remains understudied.

By gaining a more complete understanding of the genetic basis for PAS choice and the

level of PAS usage conservation in primates, it will be possible to place APA into complex models of gene regulation. It is likely that genetic variation associated with APA *(apaQTLs)* will both help to explain eQTLs outside of enhancers and promoters as well as GWAS loci not previously associated with eQTLs. Variation in APA may also contribute to how genes are differentially expressed between humans and non-human primates at both the mRNA and protein levels.

The goal of my work is to begin to characterize APA in populations of humans and non-human primates in an experimentally tractable cell type. I have ensured the methods and analysis pipelines are available for others so that they can expand this important work to a wider range of cell types and biological processes. Together these studies will contribute to the understanding of non-coding genetic variation and the molecular processes connecting genotype to phenotype.

1.4 Dissertation overview

The work I will present in this dissertation aims to characterize the role of APA in gene regulation. In chapter 2, I take a quantitative genetic approach by identifying genetic variation associated with APA in a panel of 52 human lymphoblastoid cells lines *(LCLs)*. In chapter 3, I use a comparative primate genomics approach to identify both conserved and divergent cases of APA between human and chimpanzee. In chapter 4, I describe an attempt to understand how transcription dynamics contributes to variation in both alternative splicing and APA. Finally, in chapter 5, I conclude the dissertation by placing my work in context with other recent work in the area and describe my recommendations for future directions.

CHAPTER 2

ALTERNATIVE POLYADENYLATION MEDIATES GENETIC REGULATION OF GENE EXPRESSION

2.1 Abstract[1]

Little is known about co-transcriptional or post-transcriptional regulatory mechanisms linking noncoding variation to variation in organismal traits. To begin addressing this gap, we used 3' Seq to study the impact of genetic variation on alternative polyadenylation (APA) in the nuclear and total mRNA fractions of 52 HapMap Yoruba human lymphoblastoid cell lines. We mapped 602 APA quantitative trait loci (apaQTLs) at 10% FDR, of which 152 were nuclear specific. Effect sizes at intronic apaQTLs are negatively correlated with eQTL effect sizes. These observations suggest genetic variants can decrease mRNA expression levels by increasing usage of intronic PAS. We also identified 24 apaQTLs associated with protein levels, but not mRNA expression. Finally, we found that 19% of apaQTLs can be associated with disease. Thus, our work demonstrates that APA links genetic variation to variation in gene expression, protein expression, and disease risk, and reveals uncharted modes of genetic regulation.

1. Citation for chapter: Mittleman BE, Pott S, Warland S, Zheng T, Mu Z, Kaur M, Gilad Y, and Li YI. Alternative polyadenylation mediates genetic regulation of gene expression. 2020 June 25; eLife 2020;9:e57492; DOI: 10.7554/eLife.57492

2.2 Introduction

Most genetic variants associated with complex traits are noncoding, suggesting that inter-individual variation in gene regulation plays a dominant role in determining phenotypic outcome. To investigate the function of trait-associated variants identified using genome-wide association studies (GWAS), studies have used regulatory quantitative trait loci (QTL) mapping to associate GWAS loci with variation in mRNA expression levels, DNA methylation levels, and other molecular phenotypes. Although many GWAS loci affect mRNA expression levels (i.e. are eQTLs), several recent discoveries highlight the pressing need for a better understanding of the genetic control of gene regulation, beyond that of just mRNA expression levels. For example, one recent study [34] found that the majority of autoimmune GWAS loci do not appear to affect mRNA expression levels. Two other studies observed that many genetic variants that affect protein expression levels (pQTLs) do not affect mRNA expression levels [10, 32]. Specifically, Battle and colleagues found that about half of the cis-pQTLs they identified in human LCLs (146 out of 278, 52%) did not appear to impact gene expression levels in the same lymphoblastoid cell lines (LCLs) [10]. Altogether these findings indicate that there may be unknown or understudied regulatory mechanisms that link genetic variation to complex traits, and that these mechanisms are independent of changes in the amplitude of mRNA expression levels. Moreover, even when a disease-associated variant impacts mRNA expression levels, the mechanisms by which expression is affected is often unclear. Indeed, a third of all eQTLs identified in human LCLs are not associated with variation in chromatin as measured using assays for chromatin accessibility or for modification levels of several histone marks [95]. These observations raise the possibility that understudied regulatory mechanisms mediate the effect of a substantial number of genetic variants on gene expression level.

One such understudied mechanism is alternative polyadenylation (APA). Well over half of all human protein coding genes encode multiple polyadenylation sites (PAS), resulting in

the production of diverse mRNAs with alternative termination sites [179, 114, 165]. Unlike alternative mRNA splicing, which leads to changes in splice site selection, APA leads to changes in the transcript termination site, often resulting in 3' untranslated regions (UTRs) with different lengths. As 3' UTRs are densely packed with regulatory elements that impact mRNA stability, miRNA binding, and mRNA localization (reviewed in [115, 179]), genetic control of APA may be a key mechanism by which genetic variants impact gene regulation, including mRNA expression levels, without affecting chromatin-level phenotypes such as promoter or enhancer activity. Moreover, proteins translated from different APA isoforms may differ in length and protein-protein interactions, and these differences can impact cellular phenotype. For example, globally increased usage of intronic PAS has been shown to increase risk for multiple myeloma and chronic lymphocytic leukemia [90, 168] through the translation of truncated mRNAs into truncated proteins, which impairs tumour-suppressive functions [90, 168].

To evaluate the role of APA in mediating genetic effects on gene expression and disease, we sought to identify genetic variants associated with APA on a genome-wide scale. To date, the few studies that have used genome-wide methods to identify variants associated with APA (apaQTLs) have used existing RNA-seq data to infer PAS locations and usage [92, 204, 199, 20, 107]. While using existing RNA-seq to study APA is economical, identifying PAS and estimating usage using RNA-seq are error-prone and often imprecise [62]. Furthermore, using standard RNA-seq data alone to study APA is not informative with regards to whether inter-individual differences in PAS usages are the result of variation in transcriptional termination site choice, or isoform-specific decay or export. Here, we used 3' RNA-seq (3' Seq) to measure PAS usage in steady-state mRNA collected from whole cells as well as mRNA collected from the nucleus, which is comprised of a high proportion of nascent mRNA. This design allowed us to study the effect of genetic variation on isoform PAS at multiple stages of the mRNA lifecycle. Importantly, we collected these data from a panel of human lymphoblastoid cell

lines (LCLs) that were previously profiled in great molecular detail, including measurements at the chromatin, RNA, and protein levels [42, 116, 95, 144]. Integrating the apaQTLs we identified with previously collected molecular data allowed us to study the impact of APA variation on the major steps of the gene regulatory cascade (Figure 2.1A). We use these data to show that genetic effects on APA can affect virtually all steps of gene regulation (mRNA expression level, translation rate, and protein expression level), and that such effects can impact protein expression, without affecting RNA expression.

2.3 Results

2.3.1 Alternative polyadenylation in human LCLs as defined using Nuclear and Total mRNA 3' Seq

To measure inter-individual variation in APA, we quantified PAS usage in a panel of 52 Yoruba HapMap LCLs. These same cell lines have been the subjects of multiple studies of gene regulation over the last decade [42, 116, 95, 144]. We applied 3' Seq to mRNA collected from whole cells (total mRNA fraction) of 52 LCLs and used a peak calling approach (Methods) to comprehensively identify PAS and estimate their usage (Figure 2.1B,C). Our approach obviates the need for existing annotations, which are biased towards highly expressed isoforms or isoforms expressed in well studied cell-types with higher RNA-seq coverage. In addition, to capture polyadenylated mRNA that may be under-represented or absent in the total mRNA fraction due to rapid turnover, we separately applied 3' Seq to mRNA from isolated nuclei (nuclear fraction) of the same 52 LCLs (Supplementary file 1). Because 3' Seq uses polyA priming to capture the location of polyadenylation sites and is therefore prone to internal priming at transcribed regions that are *A*-rich, we carefully filtered our data to ensure a minimal effect of mispriming on the set of PAS we considered (Method). Specifically, similar to methods previously described, we filtered both individual

reads and PAS that map to genomic regions with 70% A nucleotides or a stretch of 6 A's in the 10 nucleotides upstream [164, 178]. After quality control and filtering, we defined the usage of each PAS in a sample as the ratio of the number of reads that map to the PAS to the number of reads that map to all PAS for the same gene (Figure 2.1C) (Methods). Thus, we measure the usage of a PAS as the fraction of transcripts using that PAS over the total number of transcripts from the same gene.

We identified 41,810 nuclear PAS in 15,043 genes with at least 5% mean usage across the 52 LCL samples. We found that 67% of the protein coding genes expressed in LCLs harbor multiple PAS, suggesting that APA can impact the regulation of most genes [179, 114, 165]. Interestingly, we identified a slight negative correlation between the expression level of a gene and the number of PAS identified for the gene (Pearsons Correlation = -0.12, $p = 2.2 \times 10^{-16}$). In particular, genes with a single PAS tend to be expressed more highly than genes with multiple PAS. This observation is counter-intuitive from a statistical perspective, and it shows that, in general, out ability to detect PAS was not limited by 3' Seq coverage (Supplementary Figure 2.5, Methods). We found that the polyA binding protein motif (AATAAA), also known as the polyadenylation signal site, is the most strongly enriched motif in the 50bp regions upstream of our PAS (hypergeometric test, $p < 10^{-391}$).

We observed that PAS in the 3' UTR are more likely to have a polyadenylation signal compared with intronic PAS ($p < 10^{-16}$, difference of proportion t-test, 75.0% vs 24.8%,) (Figure 2.1D, Supplementary Figure 2.6) and that nearly half (48.3%) of all 41,810 PAS we identified are located in 3' UTRs (19.4x enrichment) [168]. Nevertheless, despite an overall depletion of PAS in introns (0.35x genome-wide levels), we found that the number of PAS in introns is notable (12,793/41,810; 30.6%) (Figure 2.1E, Supplementary Figure 2.7). While signal sites were more highly enriched near 3' UTR PAS than intronic PAS, PAS in introns show clear enrichment of polyadenylation motif 10–50 bp upstream of the cleavage site compared to background intronic sequences (24.8% vs 0.24% $p < 10^{-16}$, difference of

proportion t-test, Figure 2.1D). Thus, the recognition of intronic polyadenylation signals is a general mechanism that can result in premature termination of transcription.

We tested the hypothesis that the intronic PAS we identified correspond to truncated mRNA transcripts that escaped telescripting, whereby the U1 snRNP protects introns from premature cleavage and polyadenylation [76, 13, 128]. Because the main role of U1 snRNP is to bind and recognize 5' splice sites, the telescripting model predicts that weaker 5' splice sites can result in decreased U1 snRNP affinity for an intron and thus higher rates of early cleavage and polyadenylation. We estimated 5' splice site strength for all intron using MaxEntScore [203] and found that introns with the weakest 5' splice sites harbored more PAS than introns with stronger 5' splice sites (1.5x fold difference, first decile vs remaining deciles, hypergeometric test $p = 8.07 \times 10^{-87}$, Supplementary Figure 2.8, Methods) [180]. Moreover, we found that the top 10% of most highly used intronic PAS have weaker 5' splice sites than introns with lowly used PAS or a random set of introns (Mean MaxEntScore 6.43 vs 7.26 vs 7.48, $p = 1.4 \times 10^{-3}$, Wilcoxon rank sum test). These observations are consistent with the hypothesis that telescripting protects nascent transcripts from early cleavage and polyadenylation in introns, and that the intronic PAS we observe result from transcripts that escape telescripting [76, 13, 128].

We observed that intronic PAS have on average lower usage across individuals than PAS located in 3' UTRs (16.9% vs 46.2%). Lower usage of intronic PAS may be explained by weaker polyadenylation signals at intronic PAS compared to 3' UTR PAS or by the impact of telescripting on intronic polyadenylation. However, we hypothesized that some intronic PAS have low usage because premature polyadenylation at intronic sites can produce short-lived transcripts that are rapidly degraded and thus are under-represented in the total mRNA fraction. To test this hypothesis, we identified PAS that are used more often, or exclusively, in the nuclear fraction compared to the total mRNA fraction. By comparing PAS usage estimated in the nuclear and total mRNA fractions from all 52 individuals, we identified at

10% FDR 591 PAS in 585 genes that are used at least 20% more in the nuclear compared to the total mRNA fraction. Of these 591 PAS, 134 were found to be used by 1% or less of the transcripts in the total mRNA fraction, suggesting that these transcripts may be absent from the cytoplasm (Figure 2.1E, Supplementary Figure 2.9, Methods). Notably, we found that 387 of the nuclear-enriched PAS are intronic (Supplementary Figure 2.9), a large proportion of which (83.4% vs 43% for all PAS) are absent from a comprehensive annotation of PAS compiled from 78 human studies that used 3' Seq (Methods, Supplementary Figure 2.10) [189]. While no other study has directly measured PAS usage in nuclei, a proportion of the nuclear enriched intronic sites have been identified in a number of human tissues (up to 10%, Supplementary file 1). These findings suggest that mRNA transcripts are terminated and polyadenylated in introns at a higher frequency than generally appreciated, and that many of these isoforms escape detection from studies of total mRNA fraction owing to their rapid decay or their propensity to remain within the nucleus.

2.3.2 Genetic loci associated with variation in APA

Having established that APA can contribute to the generation of complex transcript isoforms, we sought to identify genetic loci associated with inter-individual variation in APA. We normalized each PAS usage ratio using LeafCutter [94] and tested *cis*-associations between genetic variants and PAS usage, correcting for batch and the top principal components (Methods, Supplementary Figure 2.11, 2.12, 2.13). Using 3' Seq data from the nuclear fraction, we identified 602 nuclear apaQTLs in 479 genes at 10% FDR. In the total mRNA fraction, we identified 443 apaQTLs in 353 genes at 10% FDR. For example, individuals with the C/C genotype (rs11032578) show higher usage of an intronic PAS in the *ABTB2* gene compared to individuals that are heterozygous C/T or homozygous T/T (Figure 2.2A). In both fractions, apaQTL lead SNPs are enriched near the PAS they most strongly correlate with and near the 3' ends of gene bodies (Figure 2.2B, Supplementary Figure 2.14). The

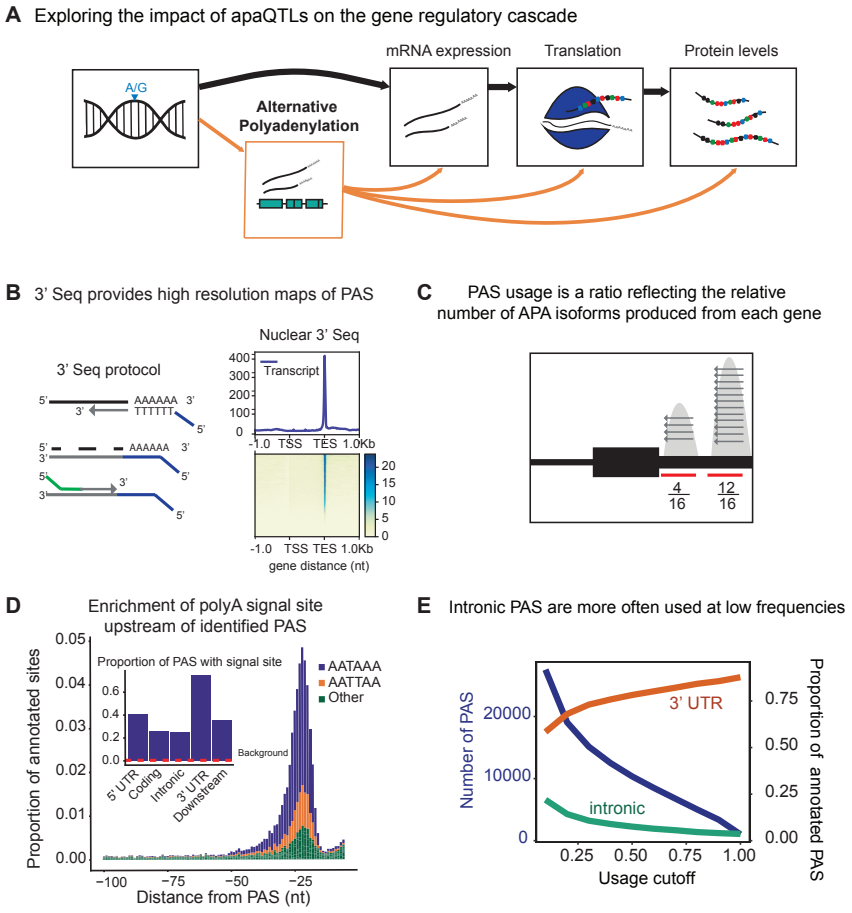


Figure 2.1: **(A)** Schematic of how genetic variants affect phenotypes by percolating through gene regulatory layers (black arrows). We aimed to understand how genetic variation can mediate gene regulation through alternative polyadenylation (orange arrows). **(B)** *(Left)* Schematic of Lexogen Reverse Quant Seq protocol for 3' Sequencing [120] *(Right)*Meta gene plot showing read coverage for five 3' Seq libraries collected from nuclei isolated from LCLs. **(C)** Representation for how PAS usage is calculated. Read count for each PAS were divided by the total number of reads at all PAS for the gene. **(D)** *(Main)* Stacked density of canonical (AATAAA, AATTAA) and other polyadenylation signal sites (AAAAAA, AAAAAG, AATACA, AATAGA, AATATA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAA) upstream of identified PAS.

Figure 2.1: (continued) *(Inset)* Proportion of PAS in different genomic regions with a polyadenylation signal site 10-50bp upstream of cleavage site. The red dotted line represents the proportion of signal site in random 40bp windows, i.e. the intronic background. **(D)** The blue line represents the number of PAS identified as the stringency of the usage cutoff increases. The orange and green lines represent the proportion of PAS in the 3' UTR and introns, respectively. The proportion of intronic PAS increases as the usage cutoff decreases, implying that a disproportionate number of intronic PAS are used at low frequencies.

proximity of the apaQTL lead SNPs to PAS may suggest that genetic variants that affect polyadenylation signal motifs drive most of the genetic effects on APA. Although we observed an enrichment of apaQTLs in signal motifs, genetic variants that alter signal motifs are unlikely to explain the majority of apaQTLs (Supplementary Figure 2.15).

Our study design provides the unique opportunity to evaluate the likely mechanisms by which genetic variation controls PAS usage. While previous studies have demonstrated that genetic variants can impact PAS usage, it has been difficult to discern whether the variation in PAS usage is primarily driven by genetic effects on cleavage and polyadenylation (Figure 2.2C, Model 1), or on the mRNA lifecycle (e.g. by impacting miRNA binding sites and decay) (Figure 2.2C, Model 2). We reasoned that if genetic effects functioned primarily by affecting post-transcriptional regulation such as decay or export, then this effect would be detectable in the total mRNA fraction, but would be smaller or undetectable in the nuclear mRNA fraction (Supplementary file 1). Interestingly, we found that only 97 apaQTLs (of 443 apaQTLs, 21.9%) identified in the total mRNA fraction were not detected in the nuclear mRNA fractions and these associations are much weaker than shared apaQTLs (Supplementary Figure 2.16). We thus suspect that we currently lack statistical power to detect most of these 97 apaQTLs in the nuclear mRNA fraction. To estimate sharing of apaQTLs across the two mRNA fractions, we used Storey's π_0 statistics and found that the vast majority of apaQTLs identified in the total mRNA fractions were estimated to also affect PAS usage in the nuclear mRNA fraction (π_1=0.87, Supplementary Figure 2.17). In addition, we found that the genetic effect sizes on PAS usage were very similar across

the two mRNA fractions ($r^2 = 0.66$; $p = 10^{-16}$, Figure 2.2D, Supplementary Figure 2.18). Altogether these observations show that most genetic variants impact PAS usage by affecting polyadenylation site choice. Supporting this notion, we found weak or no enrichment of apaQTLs in sites bound by RNA binding proteins as identified using eCLIP data from ENCODE (Supplementary file 1).

2.3.3 Impact of apaQTLs on gene expression levels

While we believe that nearly all genetic variants impact PAS usage by affecting polyadenylation site choice and not isoform-specific decay or export, this model is not incompatible with a model in which genetic variants can sometimes impact expression by affecting APA. For example, a genetic variant might increase the relative production of an isoform that is less stable, in which case total transcript levels would decrease. Therefore, next, we asked whether genetic variants could impact gene expression levels by direct effects on APA. We hypothesized that this mode of genetic regulation may be prevalent, in particular for genes with intronic PAS, because isoforms using intronic PAS are often subject to rapid decay. In this model, the genetic effect changes the relative production of isoforms with different relative stabilities rather than specifically modulating the stability of an isoform e.g. by increasing affinity for microRNA binding in the 3' UTR.

To test this hypothesis, we focused on the set of 602 apaQTLs that we identified in the nuclear mRNA fraction, representing genetic variants that impact PAS choice. Our hypothesis predicts that genetic variants that increase intronic PAS usage should decrease gene expression levels. In line with this prediction, we found a negative correlation between the genetic effect sizes for intronic PAS usage and mRNA expression levels ($p = 8.97 \times 10^{-7}$, Figure 2.3A, Supplementary Figure 2.19). Thus, our analysis suggests a widespread mechanism whereby genetic variants decrease mRNA expression levels by increasing choice of isoforms with premature PAS that are subject to rapid decay. Of interest, we found that 13

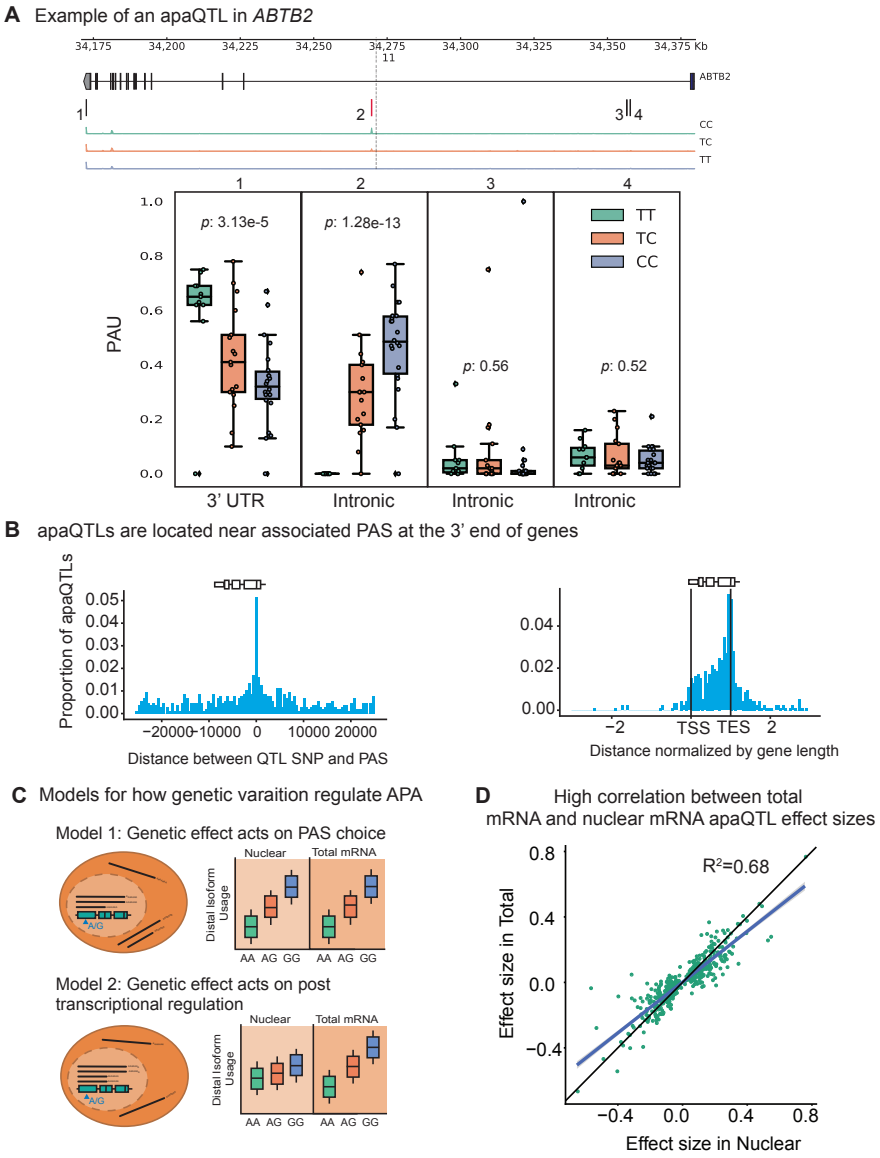


Figure 2.2: **(A)** An apaQTL in the *ABTB2* gene impact usage of an intronic PAS. *(Top)* Gene track and identified PAS. Each bar represents a potential isoform. The red bar corresponds to the isoform most strongly associated with the apaQTL. The vertical dotted line represents the position of the lead apaQTL SNP. *(Bottom)* Boxplot of polyadenylation site usage at each PAS by genotype listed according to the isoform order above. The C allele increases usage of the intronic PAS. **(B)** (*Left*) Location of the lead nuclear apaQTL SNPs relative to their corresponding PAS. (*Right*) Meta gene plot showing the distribution of apaQTL SNPs in the annotated gene body, where 0 represents the transcription start site and 1 represents the annotated transcription end site.

Figure 2.2: (continued) **(C)** Two mechanistic models for how genetic variants can affect PAS usage. (*Model 1*) Genetic variation acts directly on PAS choice. In this case, the apaQTL will be identified with similar effect sizes in both nuclear and total mRNA fractions, or smaller effect size in the total mRNA fraction. (*Model 2*) Genetic variation acts through a post transcriptional mechanism. For example, one mRNA isoform is subject to decay. In this case, the apaQTL will be identified only in the total mRNA fraction, or will be identified in the total mRNA fraction with a larger effect size than in the nuclear mRNA fraction.**(D)** Effect sizes of apaQTLs originally identified at 10% FDR in the nuclear mRNA fraction plotted against the effect sizes ascertained in the total mRNA fraction. Regression line is shown in blue and $y = x$ line is shown in black.

apaQTLs that were detected only in the nuclear fraction are also eQTLs, which highlights the importance of considering early stages of the mRNA lifecycle to uncover eQTL mechanisms.

To further investigate the contribution of APA to gene expression, we sought to understand the relationship between apaQTLs and a set of eQTLs that we previously classified as those with explained putative mechanisms, explained eQTLs (1164 eQTLs, $\sim 60\%$) or as unexplained eQTLs (801 eQTLs, $\sim 40\%$) using data from the same LCLs [95]. The eQTLs with explained putative mechanisms were associated with chromatin-level phenotypes including DNase-I hypersensitivity, histone marks, or DNA methylation, and thus are likely to be mechanistically explained by effects mediated by chromatin-level phenotypes (e.g. enhancer or promoter activity). To test whether apaQTLs might account for unexplained eQTLs, we first asked whether genes with unexplained eQTLs were more likely to also harbor apaQTLs than compared to genes with explained eQTLs. Indeed, we found a significantly higher enrichment of low p-value associations with APA for genes with unexplained eQTLs ($p = 0.01$, Figure 2.3B, Supplementary Figure 2.20) and significantly larger absolute apaQTL effect sizes for unexplained eGene compared to explained eGenes (0.35 vs. 0.3, Wilcoxon Rank sum test, $p = 6.6 \times 10^{-4}$). We also found that apaQTLs exhibited an association with chromatin states that was more similar to the unexplained eQTLs than the explained eQTLs. In particular, apaQTLs and unexplained eQTLs were more likely to lie in regions of transcription elongation or are associated with weak transcription, and less likely to lie in enhancers

or promoters than explained eQTLs (Supplementary Figure 2.21, Methods). Overall, we estimated that 17.3% of otherwise unexplained eQTLs were associated with PAS usage (see Methods). For example, an unexplained eQTL for *C10orf88* (rs7904973) colocalizes with an apaQTL associated with increased usage of an intronic PAS (Figure 2.3C). More generally, we found that eQTLs and apaQTLs colocalize for the majority of genes that had both (Methods, Supplementary file 1). These observation thus highlights APA as one important mechanism by which genetic variation impacts gene expression independent from enhancers and promoters.

2.3.4 APA mediates gene regulation independently of mRNA expression levels

Previous joint analyses of molecular QTLs suggested that functional genetic variants tend to affect gene regulation in a simple and straightforward manner: first impacting chromatin activity, then mRNA expression, and finally protein expression [95, 10]. However, because isoforms with different 3' UTRs have been shown to vary in terms of their translation efficiency, we hypothesized that apaQTLs can impact ribosome occupancy and protein expression levels without affecting mRNA expression levels [51]. To test this possibility, we asked whether apaQTLs are enriched among genes without a known eQTL, but that are associated with a ribosome occupancy QTL (riboQTL) or a protein expression QTL (pQTL). Indeed, we found that apaQTLs are enriched among genes with a ribosome QTL (rGenes; Wilcoxon rank sum test, $p = 0.01$) and genes with a pQTL (pGenes; Wilcoxon rank sum test, $p = 0.0006$) compared to genes with no molecular association (Figure 2.4A, Supplementary Figure 2.22) [95, 10]. In addition, we observed a small but significant positive correlation between individual variance in APA usage and ribosome occupancy (correlation $= 0.15$, $p < 2.2 \times 10^{-16}$, Supplementary file 1), supporting a model in which APA impacts translation efficiency.

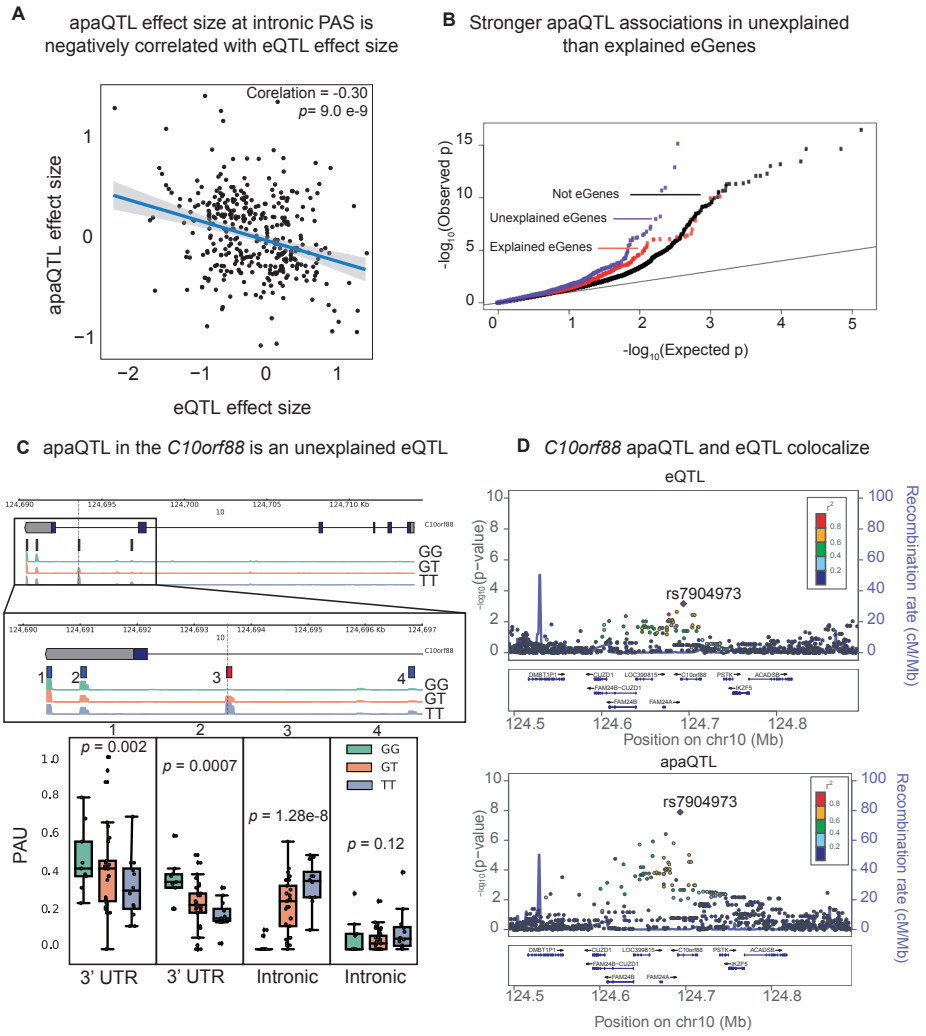


Figure 2.3: **(A)** Scatter plot of intronic apaQTL effect sizes plotted against their eQTL effect sizes shows negative correlation. **(B)** Quantile-quantile (Q-Q) plot for apaQTLs shows that apaQTLs are more highly enriched in unexplained eGenes (purple dots) compared to explained eGenes (red dots). **(C)** Example of an apaQTL that is also an unexplained eQTL for *C10orf88*. *(Top)* Gene track and identified PAS in the *C10orf88* gene. The red bar corresponds to the isoform most strongly associated with the apaQTL. The vertical dotted line represents the position of the strongest apaQTL SNP. *(Middle)* Zoomed version of track represented above. *(Bottom)* Boxplot of polyadenylation site usage at each PAS by genotype listed according to isoform order above. **(D)** *(Top)* LocusZoom plot for eQTL associations for the *C10orf88* gene. (*Bottom)* Locus zoom plot for apaQTL associations. Interestingly, the lead apaQTL and eQTL SNP, rs7904973, has been linked to increased LDL cholesterol through GWAS [83].

In total, we found 24 apaQTLs that affect protein expression, but not mRNA expression (Table 2.1). Of these, five apaQTLs were significantly associated with ribosome occupancy (Table 2.1). This finding is particularly noteworthy because nearly all genetic effects on ribosome occupancy have been proposed to be mediated by effects on mRNA expression [10]. Yet, here we provide direct evidence that APA can mediate genetic effects on ribosome occupancy without affecting mRNA expression levels. For example, the apaQTL in the *EIF2A* gene that is associated with a switched usage of two 3' UTR PAS, colocalizes with a pQTL and a ribosome occupancy QTL (Figure 2.4B, Supplementary Figure 2.23), but is not associated with *EIF2A* mRNA levels (Figure 2.4B). Interestingly, the QTL in *EIF2A* affects usage of two PAS in the same 3' UTR implying that the protein sequence encoded by the two isoforms are identical. Thus, the regulatory associations uncovered at *EIF2A* cannot simply be explained by differences in protein isoform stability. Moreover, while differences in 3' UTR are often assumed to play a regulatory function by influencing decay [115], mechanisms involving RNA decay cannot be operational in this case because steady-state mRNA expression is unchanged. Instead, differences between the two isoforms may reflect differential binding of factors that impact translation [198], or differential rates of translation re-initiation at the end of a translation cycle [158].

We identified 19 pQTLs that were associated with APA but not steady-state gene expression or ribosome occupancy levels. Two previous studies also reported the discovery of pQTLs that were not eQTLs [10, 32]. In both studies, the authors proposed that some genetic effects on protein expression levels were mediated by changes in the protein sequence or by changes in the expression level of interacting proteins, which would manifest post-translationally. Our finding reveals yet another mode of genetic regulation of protein expression level by APA (e.g. by affecting recruitment of interacting proteins). Thus, these findings provide clear evidence that APA can affect protein expression levels without affecting gene expression levels. Altogether, our findings suggest complex modes of gene regulation

independent of mRNA expression driven by variation in APA.

2.3.5 APA mediates genetic effects on complex traits

Genetic variation may impact disease risk through APA. We asked whether common variants in the regions around PAS, i.e. regions enriched for apaQTLs, are enriched in disease heritability. Using LDscore regression to estimate the heritability enrichment of 35 traits in 1kb regions centered around PAS, we found that 14 of the tested traits were significantly enriched (2.24). Of note, genetic variation around PAS was estimated to tag 15.35% of the SNP heritability for rheumatoid arthritis (7.88 fold enrichment, $p = 0.0025$). We further asked whether we could identify specific apaQTLs associated with phenotype. Indeed, 19.3% of apaQTLs (including SNPs in LD with $r^2 > 0.9$) are significantly associated with at least one trait in the UCSC GWAS catalog (Methods) [78]. For example, an apaQTL that colocalizes with the eQTLs in the C10orf88 gene (rs7904973) has been associated with increased LDL cholesterol [83], suggesting that eQTLs mediated by APA can impact organismal phenotype. Taken together, we propose that APA is a complex regulatory mechanism relevant to our understanding of how genetic variation can affect disease. Thus, comprehensive maps of apaQTLs can enhance our ability to interpret GWAS loci, particularly when the implicated variants are not eQTLs [73, 90]. For example, an apaQTL in the *ELL2* gene (rs56219066) is correlated with increased usage of an intronic PAS and is associated with risk for multiple myeloma [173]. Interestingly, multiple myeloma is among the cancer types in which widespread dysregulation of intronic APA has been documented previously [168, 90].

2.4 Discussion

Obtaining a comprehensive understanding of the mechanisms that affect gene regulation is crucial for the functional interpretation of noncoding genetic variation. Yet, existing studies that examine the role of genetic variation on APA are generally characterized by

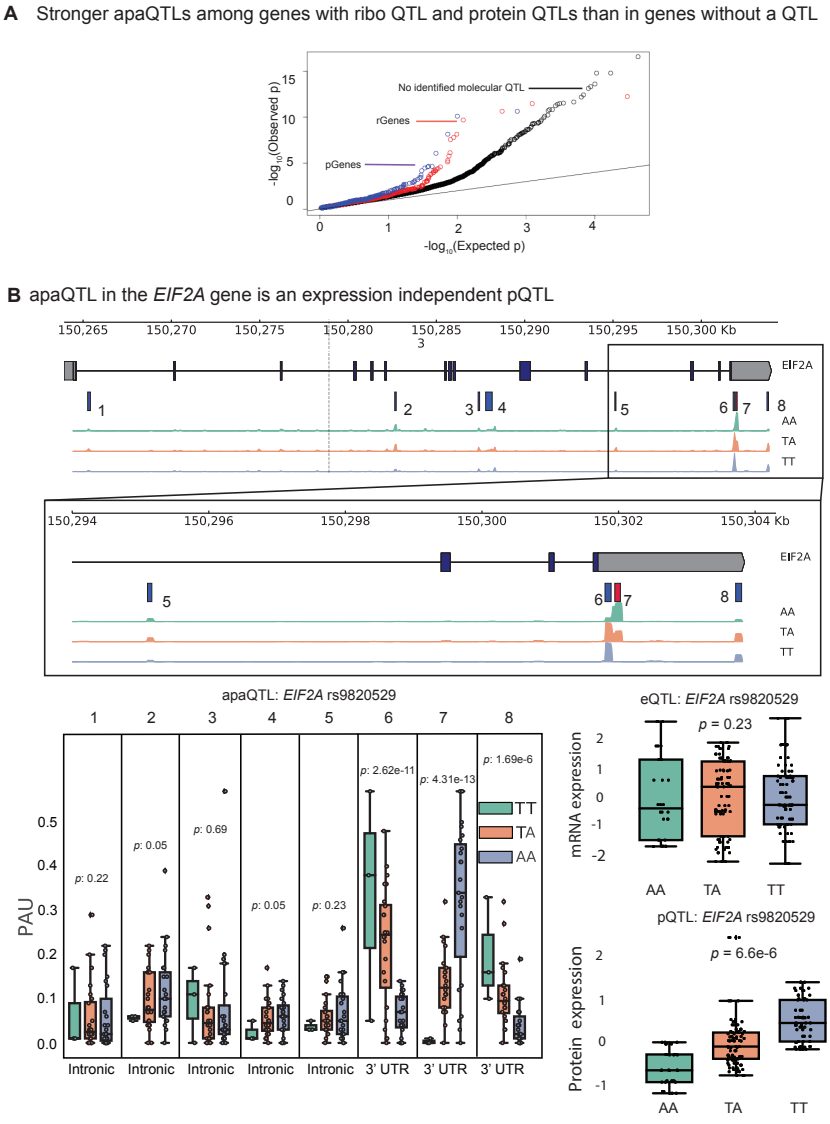


Figure 2.4: **(A)** Quantile-quantile (Q-Q) plot for apaQTLs separated by genes in previously detected rQTLS (red) and pQTLs (purple) that are not eQTLs. Black points are apaQTL genes with no pQTL, rQTL, or eQTL. **(B)***(Top)* Gene track and identified PAS in the *EIF2A* gene. The red bar corresponds to the isoform most strongly associated with the apaQTL. The vertical dotted line represents the position of the strongest apaQTL SNP. *(Middle)* Zoomed version of track represented above. *(Bottom Left)* Boxplot of polyadenylation site usage at each PAS by genotype listed according to isoform order above. *(Top Right)* Boxplot showing normalized mRNA expression for *EIF2A* by genotype at the apaQTL SNP (rs9820529). *(Bottom Right)* Boxplot showing normalized protein expression for *EIF2A* by genotype at the apaQTL SNP (rs9820529).

two important shortcomings. Firstly, the study of inter-individual variation in PAS usage have been mostly restricted to APA in the 3' UTRs [92, 204, 199], leaving genetic variants that impact PAS usage in other regions, e.g. intronic PAS, understudied. Secondly, nearly all existing studies use standard RNA-seq to estimate PAS usage, which not only limits the accuracy of usage quantification, but also makes it difficult to disentangle the contribution of co-transcriptional mechanisms to APA regulation from post-transcriptional mechanisms such as isoform-specific decay. In this study, we overcome these shortcomings by applying 3' Seq to total and nuclear mRNA fractions separately to directly measure PAS usage including that of PAS in intronic regions.

It is worthwhile to note here that despite the many advantages of using 3' Seq to identify and quantity APA isoforms, 3' Seq experiments are known to be susceptible to mispriming, which occurs when polydT primers designed to recognize the polyA-tail of transcripts anneal to adenosine stretches within a transcript, thus introducing false positive polyadenylation sites. While we used stringent criteria to reduce the effect of mispriming, we found that a small proportion of PAS used in this study may be the result of mispriming. In particular, we found an enrichment of adenosine nucleotides at a subset of intronic PAS which were discovered in our study and not previously annotated, suggesting that 10–20% of unannotated intronic PAS may be false positives (Supplementary file 1). To ensure that these false positive PAS do not affect the validity of our analyses, we performed the main analyses presented in this study after removing unannotated intronic PAS and found that our conclusion were robust to the small number of potential false positive intronic PAS (Supplementary file 1) [189].

By collecting data from both total and nuclear mRNA factions, we were able to study the effects of genetic variation on polyadenylation at multiple stages of the mRNA lifecycle, and to distinguish putative regulatory mechanisms by noting the stages at which the genetic effects on APA were observed. For example, genetic variants can impact steady-state isoform

ratio either co-transcriptionally by affecting PAS choice during transcription (Figure 2.2C top), or post-transcriptionally by affecting binding of miRNAs or RNA-binding proteins and consequently isoform decay (Figure 2.2C bottom). We found that the vast majority of genetic variants that affect PAS usage ratio in total mRNA fraction, were also found to have similar effect sizes on PAS usage ratio in the nucleus. This observation implies that inter-individual variation in steady-state APA levels can generally be explained by variation in co-transcriptional mRNA processing, or mRNA processing that occur soon after transcription.

There are several co-transcriptional mechanisms that may result in variation in PAS usage. For example, previous reports have suggested that variation in the polyadenylation signal site may cause variation in PAS usage. While we found that this was the case for a small number of examples, disruption of canonical signal motifs does not appear to be a major mechanism for generating apaQTLs, an observation that is also supported by a recent study on APA in GTEx data (Supplementary Figure 2.15) [92]. Other possible co-transcriptional mechanisms involved in PAS choice include competition between the spliceosome and polyadenylation factors for example mediated by the spliceosomal RNA U1 [128], and RNAP II pausing [53]. Indeed, recent studies have reported that sequence and chromatin context can pause or slow down RNAP II elongation across the gene body [112], suggesting that variation in RNAPII pausing may impact PAS choice [53]. For example, in Drosophila melanogaster, paused RNAPII promotes the recruitment of ELAV on the pre-mRNA, which prevents usage of a proximal PAS [130]. In addition, Liu et al. observed a tissue-specific shift toward usage of proximal PAS sites in Drosophila melanogaster mutant for a slow elongation form of RNAPII [102]. These findings further suggest that variants affecting RNAPII elongation rate could underlie the genetic effects on PAS usage we detected in this study.

Although our data suggest that apaQTLs do not generally impact rates of mRNA decay, e.g. by affecting miRNA or RBP binding motifs, we found clear evidence that apaQTLs may

promote polyadenylation site choices that result in the production of isoforms with different rates of decay. For example, we observed that genetic variants that increase the usage of isoforms ending at intronic PAS tend to be associated with lower levels of gene expression. This observation is consistent with reports that isoforms with premature polyadenylation are often substrates for nonsense mediated decay or nonstop decay [179, 187]. More generally, our results suggest that apaQTLs can affect gene expression levels post-transcriptionally by impacting the production of isoforms with varying levels of stability. Importantly, our study highlights APA as an eQTL mechanism independent of promoters and enhancers.

While the effect of genetic variants on gene regulation is generally assumed to move linearly from chromatin, to mRNA, to protein level, our study reveals several complex modes of genetic regulation for both gene expression and protein expression levels by APA. Although we were unable to study the genome-wide effects of APA on protein expression owing to a scarcity of protein-level data, we identified several apaQTLs that affect protein, but not gene expression levels. These results strongly suggest that APA can affect protein expression levels without affecting gene expression levels, because our power to detect genetic effects on gene expression levels far exceeds that to detect genetic effects on protein expression levels. Furthermore, some of these pQTLs were associated with ribosomal occupancy and some were not, which implies multiple pathways by which genetic variants can impact protein expression levels through APA.

In conclusion, there are many pathways through which genetic variants can impact gene regulation and, consequently, organismal phenotypes. While many studies have demonstrated the importance of gene expression regulation through promoters or enhancers, very few studies have focused on co- or post-transcriptional gene regulation. Our study shows that co- and post-transcriptional processes such as APA can mediate the effects of a substantial number of genetic variants on mRNA expression levels, protein expression levels, and risk for complex diseases.

2.5 Methods

2.5.1 Cell Culture

We cultured 54 Epstein-Barr virus transformed LCLs under identical conditions at 37 C and 5% CO2. These LCLs were derived from Yoruba individuals originally collected as part of the HapMap project [69].The sampleIDs and Research Resource Identifiers (RRIDs) can be found in online version of paper (see chapter citation). Details for each cell line are found in Supplementary file 3. We grew cells in a glutamine depleted RPMI [RPMI 1640 1X from Corning (15-040-CM)], completed with 15% FBS, 2mM GlutaMAX (from gibco (35050-061), 100 IU/ml Penicillin, and 100 ug/ml Streptomycin. After passaging them 3 times the lines were maintained at a concentration of 1×10^6 cells per mL. In preparation for extraction, we allowed the cells to grow until a concentration of 1×10^6 cells per mL was reached and then proceeded to extraction.

2.5.2 Collection and RNA extraction

We collected 30 million cells from each line and divided them into two 15 million cell aliquots. We spun the cells down at 500 RPM at 4C for 2 min, and then washed the pellets with phosphate-buffered saline (PBS) and spun down again. After this we aspirated the PBS, leaving the cell pellet. All washing steps occurred on ice or in cooled centrifuges. At this point every cell line had two separate pellets each from an input of 15 million cells. From each line we took one of these pellets for nuclear isolation. We then carried out nuclear isolation using the nuclear isolation steps outlined by [111]. Once we washed and spun down the pellets in the nuclei wash buffer, we resuspended them in 700 ul of the QIAzol lysis reagent (Qiagen). We extracted both RNA cell pellets from the same line in the same batch using the miRNeasy kit (Qiagen) according to manufacture instructions, including the DNase step to remove potentially contaminated genomic DNA. Details for the collection such as cell

viability and cell concentration at time of collection are found in Table 2.2. We checked the quality of the collected RNA using a nanodrop. RNA concentrations and absorbance levels from the collection are in Table 2.2.

In order to verify fraction separation, we completed the Mayer and Churchman protocol to isolate chromatin and collected cell lysates for each step in the fractionation [111]. We performed western blots against both GAPDH (GAPDH antibody (6C5) Life Technologies AM4300) and the Carboxyl Terminal Domain of Pol-II (CTD) (Pol II CTD Ser5-P antibody, Active Motif, 61085). We ran each lysate on Mini-protean TGX precast gels (bioRad 456-1093) after digesting any remaining DNA molecules from the nuclear isolate with benzonase nuclease. We used Goat anti-Mouse IgG (H+L) (Invitrogen 32430) as a secondary antibody for the GAPDH antibody and Goat anti-Rat IgG (H +L) (Invitrogen 31470) as a secondary antibody for the CTD antibody. We diluted all antibodies in a 1:1000 dilution with blocking solution made from dry milk (LabScientific Lot 1267N Cat M0841). We show GAPDH isolated in the cytoplasm and CTD to the chromatin fraction (Supplementary Figure 2.25).

2.5.3 3' Sequencing library generation

We generated 108 single-end RNA 3'sequencing libraries from the total and nuclear RNA extract using the QuantSeq 3 ' mRNA-Seq Library Prep Kit [120] as directed by the manufacturer. We used 5ng of each sample as input. We submitted the libraries for sequencing on the Illumina NextSeq5000 at the University of Chicago Genomics Core facility using single end 50bp sequencing.

2.5.4 3' Sequencing data processing

We mapped 3' Seq reads to hg19 [35] using STAR RNA-seq aligner [45] using default settings with the WASP mode to filter out reads mapping with allelic bias [185]. Similar to previously published 3' Seq methods, we accounted for internal priming by filtering reads preceded by

6 Ts in a row or 7 of 10 Ts in the 10 bases directly upstream of the mapping position in the reference genome [178, 164, 12]. We verified the individual identity of all bam files using VerifyBamID [75]. Due to low confidence in the identity of 2 individuals, they were removed from all analysis. Raw read and mapped read statistics after accounting for internal priming can be found in Table 2.2 (Supplementary Figure 2.26, 2.27 2.28, 2.29).

2.5.5 Identification and characterization of PAS

We merged all mapped reads and called peaks using an inclusive method, identifying all regions of the genome with non-zero read counts in 90% percent of libraries and an average read count of greater than 2 counts. This resulted in 138,181 peaks. We assigned each of these peaks to a genic location according to NCBI Refseq annotations for 5' UTRS, 3' UTRs, exons, introns, and regions 5kb downstream of annotated genes downloaded from the UCSC table browser [78]. When a region mapped to multiple genes we used a hierarchical model, similar to the method used by Lin et al. [101] to assign the peak to a gene annotations. Our method prioritizes annotations in the following order: 3' UTRs, 5kb downstream of genes, exons, 5' UTRs, and introns. To further verify absence of PAS detected as a result of internal priming we removed PAS with 6A's or 70% As in the 15 basepairs downstream of the site. We next utilized a gene level noise filter to account for non-uniform read coverage across the genome. We created a usage score for each PAS based on of the number of reads mapping to the PAS over the number of reads mapping to any PAS associated with the same gene. We filtered out peaks with a mean usage of less than 5% in both the total and nuclear libraries. After this filter, we were left with 35,032 PAS in the total mRNA fraction and 39,164 PAS in the nuclear fraction. The merged set with PAS from both fractions used for PAS QC is available on GEO and has 41,810 PAS. We compared our set of PAS to the human PolyADB release 3.2 annotation [189](Supplementary Figure 2.10). We explored the relationship between number of PAS detected and gene expression using TPM

estimates from YRI LCLs after removing very lowly expressed genes (less than 1 TPM) [89]. We calculated the 5' splice site strength using the MaxEntScore tool, for each of the introns in our annotation [203]. We binned the introns by decile according to the scores and evaluated the distribution of the introns containing PAS. We also used the scores for the introns containing PAS to investigate the relationship between PAS usage and 5' splice site strength.

2.5.6 PAS Signal site enrichment and locations

We used the Homer findMotifsGenome.pl script with the -size -300,100 option to identify binding motifs in the 50bp upstream of each PAS [63]. As a background, we used genome shuffle to randomly chose the same number of 50bp regions. To explore the location of the signal site relative to the PAS (most 3' end of each identified peak), we determined the relative position of previously described potential signal sites to this position [12]. We then extended each PAS 100bp upstream and identified the starting position of each of the 12 PAS signal site variations identified by Beaudoing et al. without allowing for sequence mismatch [12].

2.5.7 Differential Isoform analysis

We mapped 3' Seq reads to all PAS peaks with mean coverage of 5% in the total or nuclear fraction libraries. This results in 41,813 annotated sites. We assigned reads to PAS using the featureCounts tool with the -O flag to assign reads to all overlapping features [98]. We ran the leafcutter_ds.R script on chromosomes 1-22 separately using the cellular fraction label as the sample group identifier [94]. This analysis tests 9790 genes and resulted in 8227 genes with significant (FDR 10%) isoform level differences between the total and nuclear cellular fraction. We called differentially used PAS as sites with a Δ polyadenylation site usage (Δ PAU) greater than 0.2 or less than -0.2. In our analysis a positive ΔPAU corresponds to

increase usage in the total cellular fraction while a negative Δ PAU corresponds to increased usage in the nuclear fraction.

2.5.8 apaQTL calling in both fractions

We used the leafcutter prepare_phenotype_table.py script with default settings to normalize the PAS usage ratios across individuals within each fraction. This method also outputs the top principal components (PCs) of the data to use as covariates. We plotted the proportion of variation explained by each PC in order to identify the number of PCs to include in the analysis (Figure 2.13). We included the top 4 PCs as well as the library preparation batch as the covariates. We plotted the proportion of variance explained by a number of cofactors in each of the top 10 PCs. (Supplementary Figure 2.13) The top four PCs correlate most strongly with the cell count at collection (Supplementary Figure 2.13). We used the same genotypes from Li et al. 2016[95], available at http://eqtl.uchicago.edu/jointLCL/genotypesYRI.gen.txt.gz [95]. We removed individual NA19092 due to lack of genotype information in this file, bringing our sample size to 51 individuals for this part of the analysis. Only SNPs with a MAF $> 5\%$ in our sample were included. We used FastQTL to map apaQTLs in cis (25kb on either side) with 1000 permutations to select the top SNP-PAS association [132]. We called apaQTLs in each fraction as variants passing 10% FDR (Benjamini-Hockberg) after permutations. In order to plot interpretable effect sizes for each association we computed nominal PAS:SNP associations for the pre-normalized PAS ratios.

2.5.9 Association of apaQTLs with chromatin states

We downloaded the GM12878 chromatin HMM annotations for Hg19 from the UCSC table browser [78]. We overlapped the eQTLs identified and published in Li et al. 2016[95] as well as the total and nuclear fraction apaQTLs with these categories. We calculated 95% confidence intervals for each measurement by sampling the number of QTLs in the set with

replacement 1000 times (Supplementary Figure 2.21).

2.5.10 apaQTL overlap with eQTLs

We obtained the set of explained and unexplained eQTLs from Li et al. 2016 [95]. In order to test whether genes with an unexplained eQTL are more likely to be explained by variation in APA, we separated the permuted apaQTL association (top snp per PAS) into three categories: unexplained eGene, explained eGene, non eGenes. We tested for significant enrichment of apaQTLs in each category using one-sided Wilcoxon rank sum tests. In order to test if each explained and unexplained eQTLs described in Li et al. 2016[95] overlaps with an apaQTL, we extracted the nominal associations for each eQTL gene-SNP pair from the apaQTL data in both fractions. In order to account for multiple PAS associations for each pair, we selected the most significant p-value and used a Bonferroni correction to account for the number of PAS tested in the gene. We consider an eQTL as explained by an apaQTL if the corrected p-value is less than 0.05 but report the values for a range of cutoffs in Supplementary Figure 2.30. We performed colocalization with the R coloc package [188]. The Bayes Factor colocalization method reports Bayes Factors for 4 alternative hypotheses. PP0: No association with either trait, PP1: No association with trait 1, PP2: No association with trait 2, PP3: Association with trait 1 and trait 2, two independent SNPs, and PP4: Association with trait 1 and trait 2, one shared SNP. If causal SNPs for an apaQTL and an eQTL is the same SNP, then PP4 is expected to be large (greater than 0.5). We accounted for incomplete power using the method described in Ongen et al. (Supplementary file 1) [131].

2.5.11 apaQTLs overlap with ribosome specific and protein specific QTLs

The list of protein specific QTL genes can be found in the supplementary information from Battle et al. 2015[10]. In order to show that genes with an eQTL and protein specific

QTLs are likely to be associated with APA, we separated the permuted apaQTL association (top snp per PAS) into three categories: eGene, pGene, or neither pGene nor eGene. We performed the same analysis with rGenes, eGenes, and neither rGenes nor eGenes. We tested for significant enrichment with one sided Wilcoxon rank sum tests (Figure 2.4A, Supplementary Figure 2.22).

2.5.12 Identification of molecular QTL associations

We sought to test if SNPs identified as apaQTLs are significantly associated with other molecular phenotypes previously tested in the same panel of LCLs. We tested for associations between the genotypes used in this study and each gene for each phenotype with fastqtl using the top 5 PCs calculated in Li et al. 2016 as covariates [95]. We used normalized RNA expression, RiboSeq values, and protein levels, published in Li et al. 2016 [95].

2.5.13 PAS heritability estimates and apaQTL overlap with GWAS Catalog

We downloaded GWAS summary statistics from both Astle et al. and Okada et al. [5, 129] We augmented our PAS sites by 500bp on either side and ran LD score regression using methods described in Bulik-Sullivan et al. [25] We downloaded the CRCh37hg19 GWAS catalog for UCSC table browser [78]. We identified SNPs in LD with the nuclear apaQTLs using the LDproxy tool from LDlink with YRI as the population [105]. We filtered all results to SNPs with an r2 greater than 0.9. We overlapped the full set with the GWAS catalog using pybedtools.

2.5.14 Data and code availability

Fastq files and PAS annotations are available at GEO under accession GSE138197 `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138197`. All reproducible scripts and software versions can be found at `https://brimittleman.github.io/apaQTL/` A versioned

release of the github is available through Zenodo with doi:10.5281/zenodo.3905372 `https://zenodo.org/record/3905372#.XvKD4S2ZMXp`

2.6 Acknowledgments

We thank N. Gonzalez, J.P. Staley, M.C. Ward for comments on the manuscript. **Funding**: This work was supported by the US National Institutes of Health (R01GM130738 to Y.I.L). B.E.M. supported by T32 GM09197 to the University of Chicago and F31HL149259 to B.E.M. from National Heart, Lung, And Blood Institute of the National Institutes of Health. SP was in part supported by the National Center for Advancing Translational Sciences of the NIH (K12 HL119995). This work was completed in part with resources provided by the University of Chicago Research Computing Center.

2.7 Author Contributions

Y.I.L. conceived of the project. B.E.M, S.W. and S.P. performed the experiments. B.E.M analyzed the data with help from Y.I.L, S.P., T.Z., Z.M. and M.K. B.E.M. drafted the manuscript with input from Y.G., Y.I.L, and S.P. S.P., Y.G. and Y.I.L. supervised this project.

2.8 Supplementary Information

2.8.1 Supplementary Figures

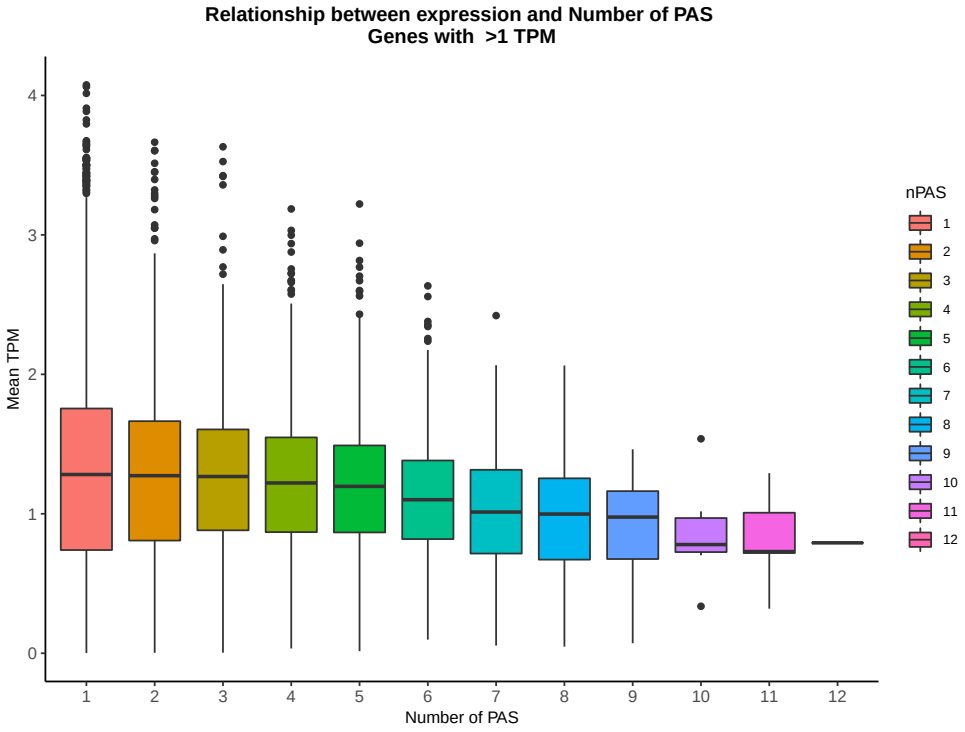


Figure 2.5: **Relationship between Number of PAS and gene expression** Relationship between number of PAS identified in our study and gene expression levels (TPM) as measured from GEUVADIS YRI LCLs [89]. Genes with mean TPM < 1 across individuals were considered not expressed and thus were removed for this analysis.

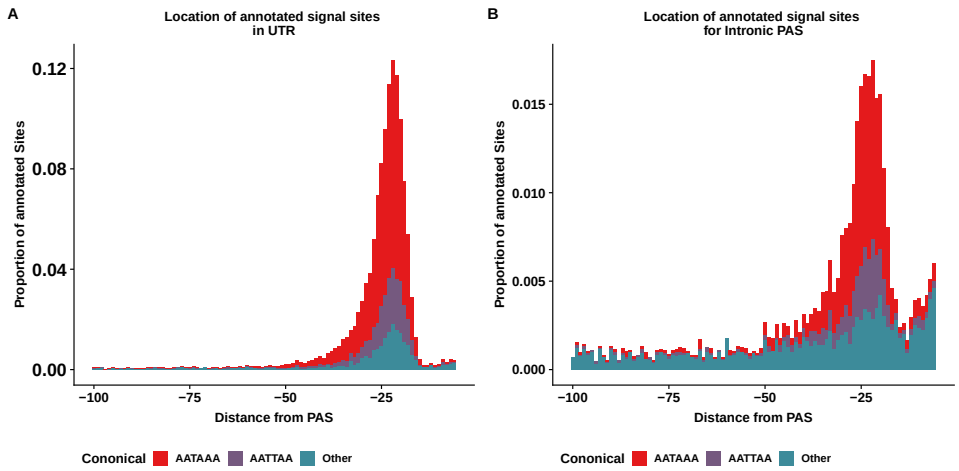


Figure 2.6: **Distribution of signal sites upstream of PAS. Supplement to Figure 2.1D (A)** Stacked density plot showing the signal site distribution for PAS in 3' UTR. Other signal sites are AAAAAA, AAAAAG, AATACA, AATAGA, AATATA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAAA. **(B)** Stacked density plot showing the signal site distribution for PAS in introns. Other signal sites are AAAAAA, AAAAAG, AATACA, AATAGA, AATATA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAAA.

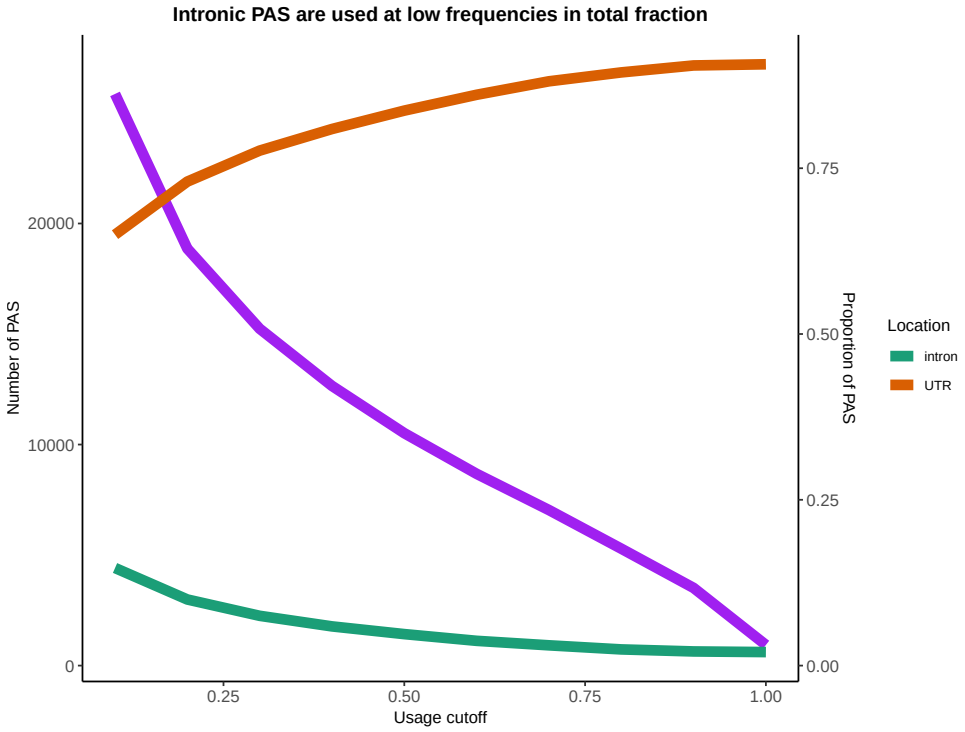


Figure 2.7: **Proportion of PAS in 3' UTRs and introns as predicted from total 3' Seq. Additional figures corresponding to Figure 2.1E.** Number of PAS identified with usage larger than the usage cutoff (x axis) in the total mRNA fraction (purple). Proportion of PAS in introns when PAS are filtered by total usage (green). Proportion of PAS in 3' UTRs when PAS are filtered by total usage (orange).

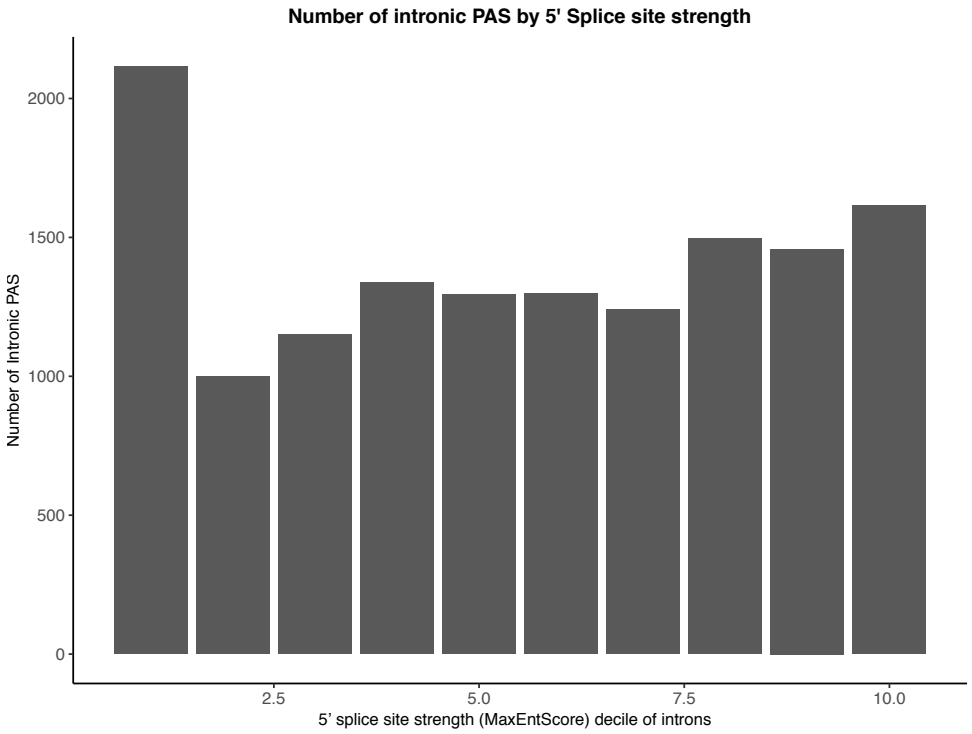


Figure 2.8: **Intronic PAS 5' Splice site strength** Intronic PAS are enriched in introns with the weakest 5' splice sites. Splice site strengths for all introns were calculated using MaxEntScore [203]

.

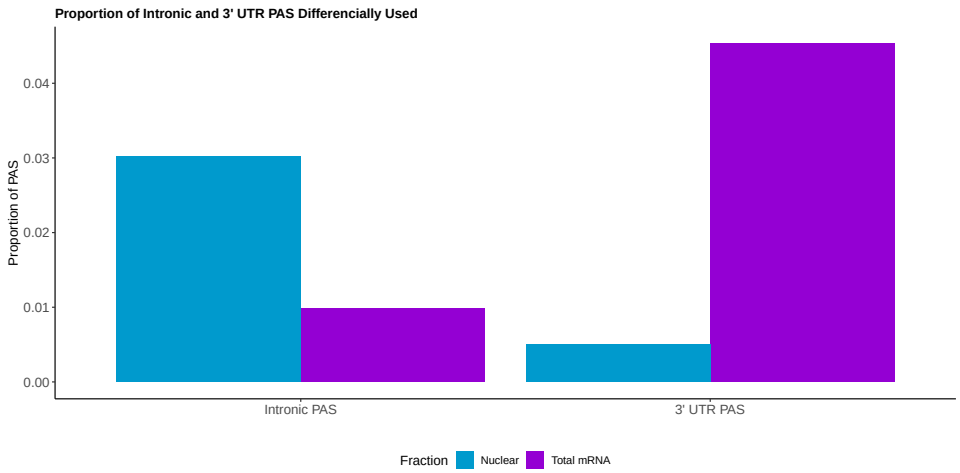


Figure 2.9: **Location of PAS differentially used** We adapted LeafCutter to identify genes with significant differential usage of PAS between the total and nuclear fraction. The majority of PAS preferentially used in the nuclear fraction are intronic, whereas the majority of PAS preferentially used in the total fraction lie in the 3' UTR.

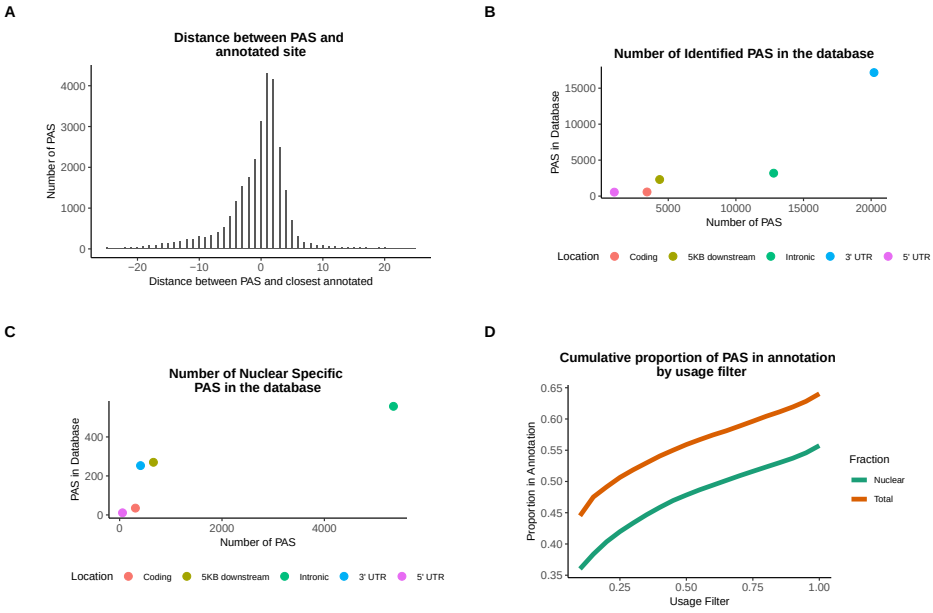


Figure 2.10: **Comparison of our 3' Seq PAS to previous PAS annotations (A)** Distribution of distance between PAS and closest annotated site in the annotation database (PolyA_DB release 3.2) [189]. **(B)** Scatter plot showing the number of PAS we identified in our study (X axis) versus the number of PAS in the PolyA database (Y axis) separated by genomic location (colors). **(C)** Scatter plot showing the number of nuclear-specific PAS we identified in our study versus the number of PAS in the PolyA database separated by genomic location (colors). The vast majority of nuclear-specific PAS are intronic. **(D)** Proportion of PAS present in the PolyA database by usage in nuclear (green) or total (orange) mRNA fraction.

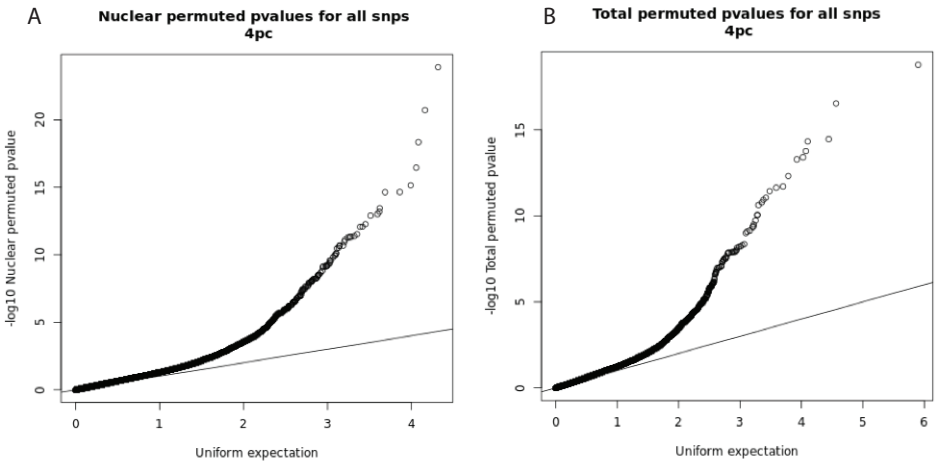


Figure 2.11: **Q-Q plots for apaQTLs (A)** Q-Q plot for nuclear apaQTLs, plotting adjusted p-values of the top SNP PAS associations. **(B)** Q-Q plot for total apaQTLs, plotting adjusted p-values of the top SNP PAS associations.

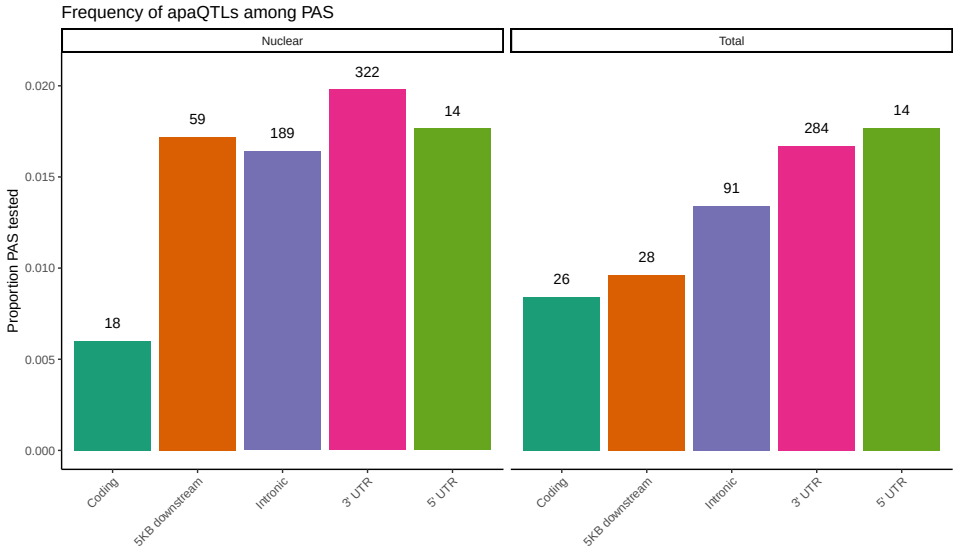


Figure 2.12: **Proportion of PAS tested with an apaQTL** Proportion of PAS in different genomic locations with a significant apaQTL. The numbers above each bar represent the number of identified apaQTL for each location.

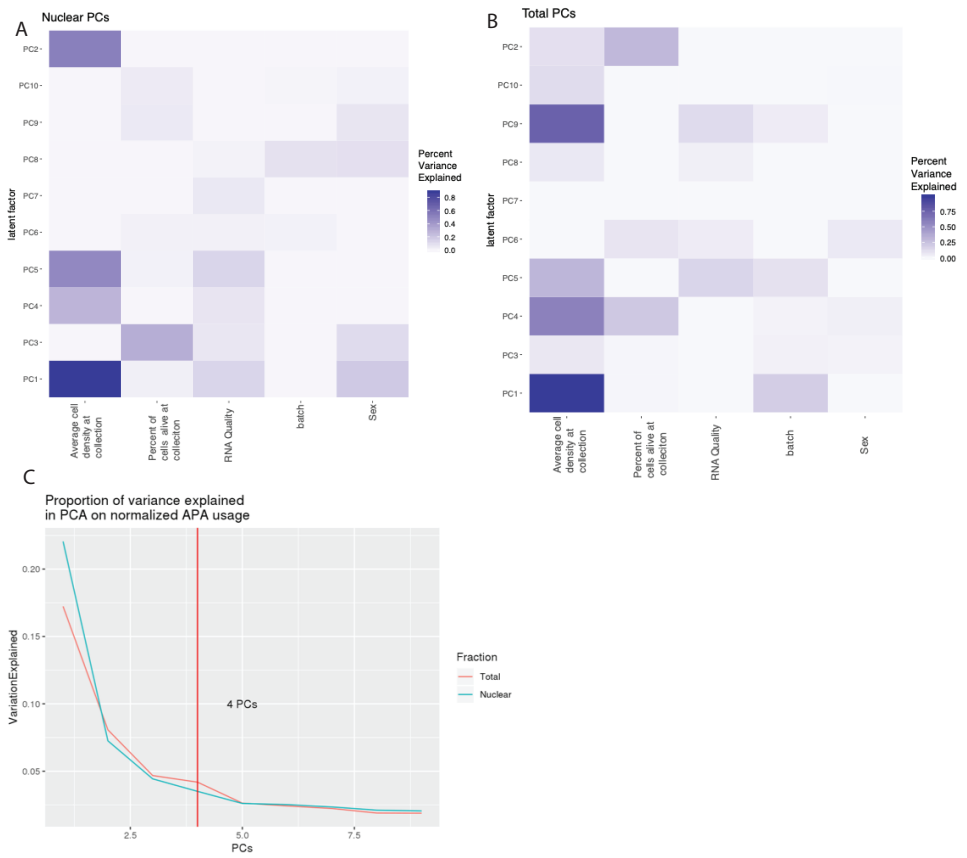


Figure 2.13: **Analysis of the PCs of APA usage (A)** Proportion of variance explained in the 10 first PCs by experimental variables in nuclear APA usage. We used a linear model to look at correlation between PC and each covariate. **(B)** Proportion of variance explained in the 10 first PCs by experimental variables in total APA usage. We used a linear model to look at correlation between PC and each covariate. **(C)** Proportion of variance explained by each PC in APA usage. Vertical line represents the number of PCs used as covariates in our apaQTL analysis.

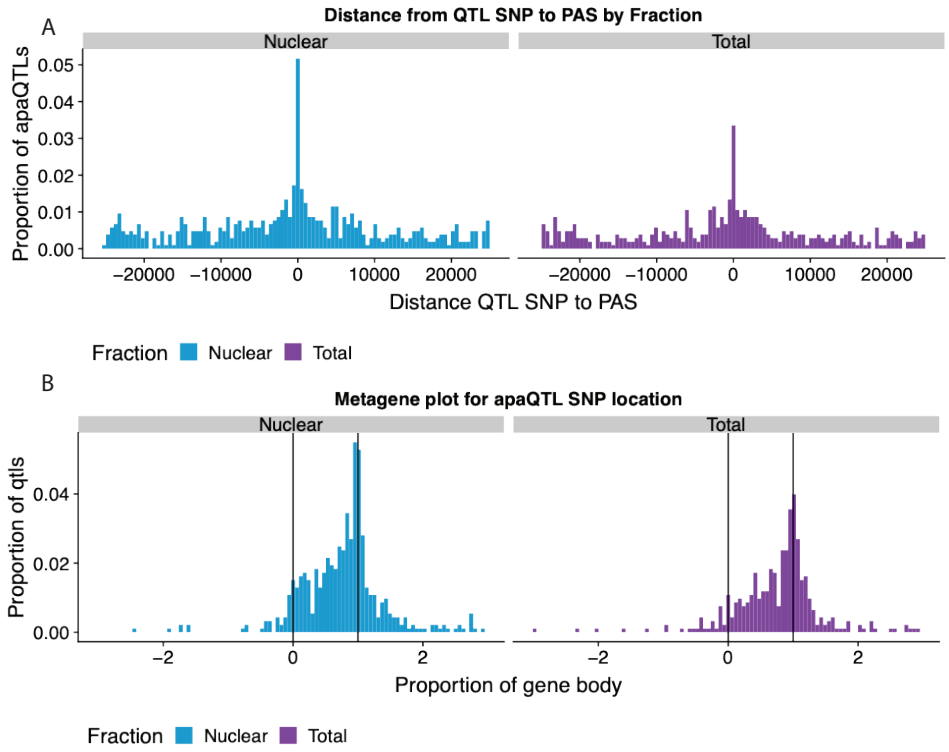


Figure 2.14: **apaQTLs in both fractions are associated with PAS near SNP and at the transcription end site. Supplement to Figure 2.2B and 2.2C. (A)** Histogram showing the distribution of the distance between lead apaQTL SNP and the PAS, separated by mRNA fraction. **(B)** Histogram showing the distribution of the distance between lead apaQTL SNP and gene features, where 0 represents annotated TSS and 1 represents annotated TES, separated by mRNA fraction.

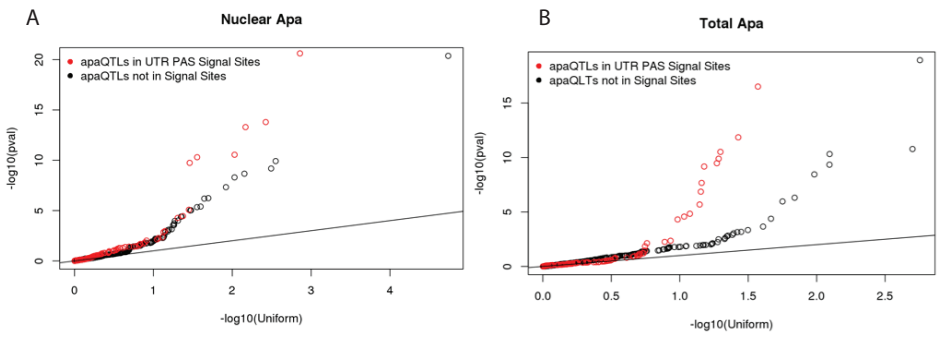


Figure 2.15: **Signal site disruption (A)** Q-Q plot showing the nuclear apaQTL p-values for SNP in signal sites upstream of 3' UTR PAS compared to matched SNPs (equal distance) upstream of a set of 3' UTR PAS without identified signal sites. **(B)** Similar to panel A, but for total apaQTLs.

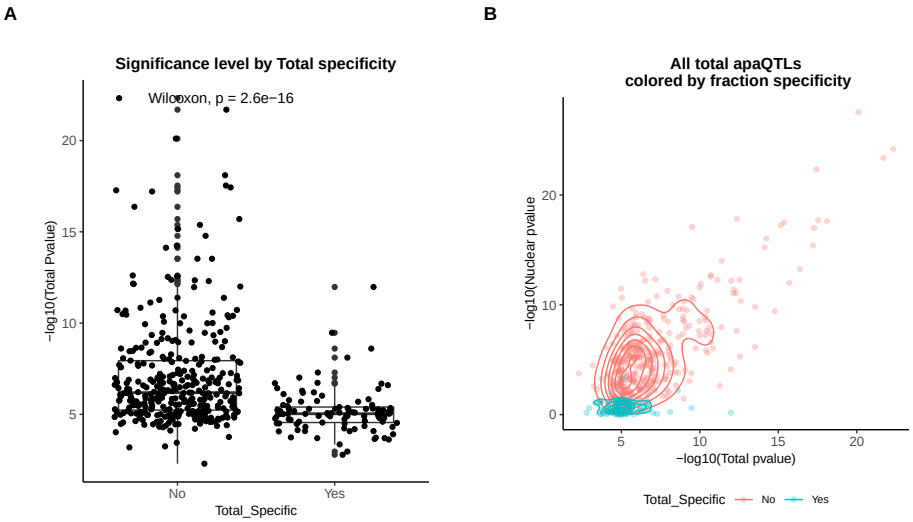


Figure 2.16: **Total mRNA specific apaQTLs show weaker association than do shared apaQTLs (A)** Boxplot showing the -log10(p-value) of the nominal total apaQTL associations separated by whether the association is also identified in the nuclear mRNA fraction. ApaQTLs that are total-specific have significantly weaker associations. **(B)** Scatter plot showing the relationship between the -log10(p-valued) of the apaQTL associations in both mRNA fractions for total mRNA apaQTLs. Dots and densities are colored by whether the apaQTL is total-specific or shared. Total-specific apaQTLs are likely not detected in the nuclear fraction due to a lack of power.

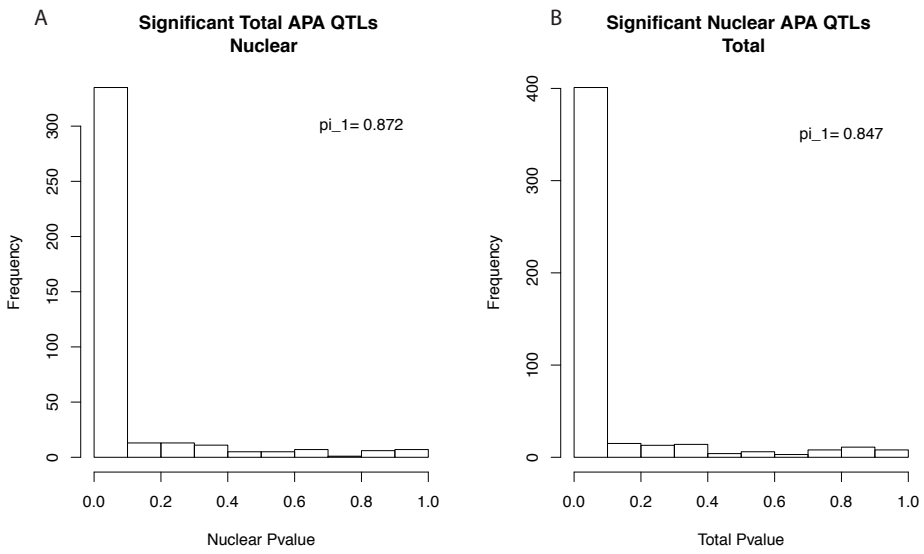


Figure 2.17: **apaQTL sharing between fractions (A)** Histogram showing the P-value distribution of the apaQTL associations between the lead total apaQTL SNP and the corresponding PAS ascertained using our 3'-Seq data from the nuclear mRNA fraction. Values were calculated based on PAS tested in both fractions (403 of 443). Results are robust to using all PAS ($pi_1 = 0.842$) **(B)** Histogram showing the P-value distribution of the apaQTL associations between the lead nuclear apaQTL SNP and the corresponding PAS ascertained using our 3'-Seq data from the total mRNA fraction. Values calculated based on PAS tested in both fractions. (483 of 602) Results are robust to using all PAS ($pi_1 = 0.825$)

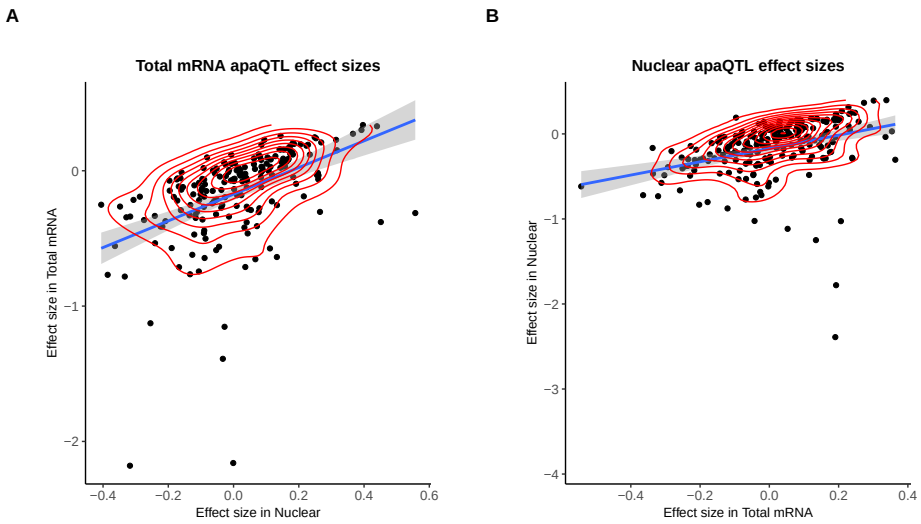


Figure 2.18: **Correlation of effect sizes for apaQTLs discovered in total and nuclear mRNA fractions (A)** Normalized effect sizes ascertained in total mRNA and nuclear fraction of total apaQTLs tested in both fractions. **(B)** Normalized effect sizes ascertained in total mRNA and nuclear fraction for nuclear apaQTLs tested in both fractions.

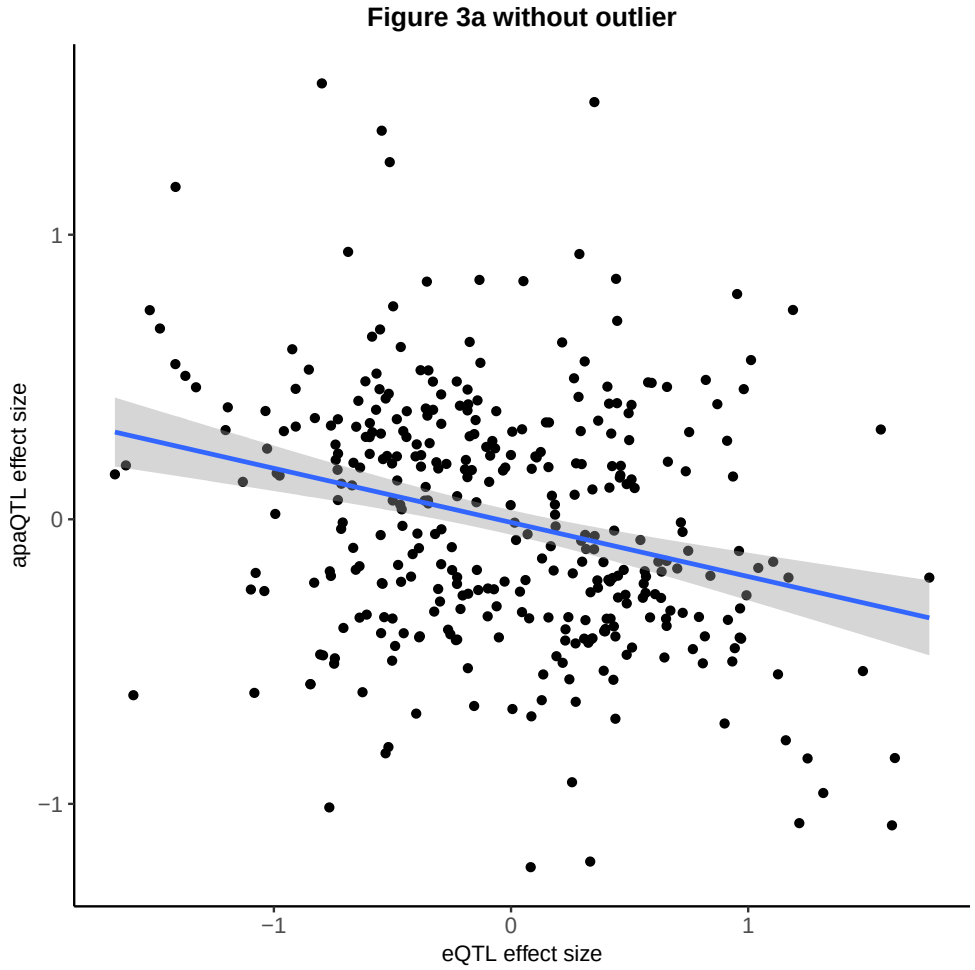


Figure 2.19: **Figure 2.3A without outlier SNP** Scatter plot showing the relationship between intronic nuclear apaQTL effect size and eQTL effect size after removing outlier SNPs (Filtered for SNPs with eQTL effect size < -2.0).

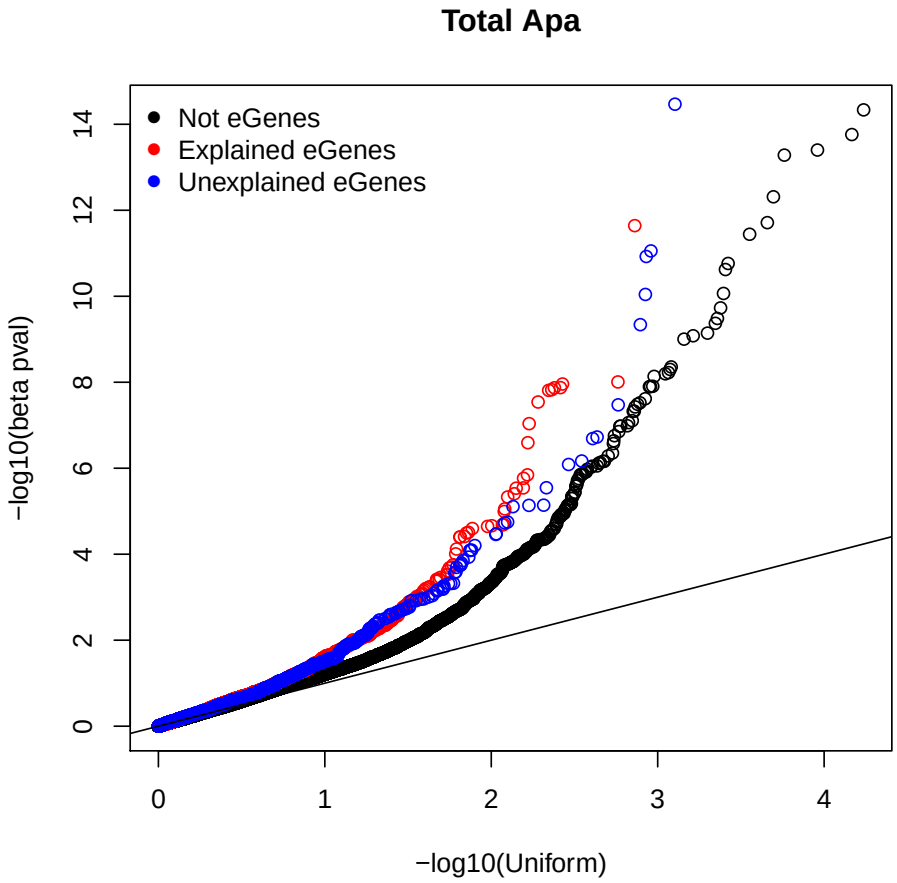


Figure 2.20: **Overlap between apaQTLs in total fraction and eQTLs, supplement to Figure 2.3B** QQ-plot showing the total apaQTL (adjusted) p-values separated by whether the gene harbors an explained (red) or unexplained (blue) eQTLs. We observe an enrichment for low apaQTL association p-values in genes with eQTLs compared to all tested genes (black).

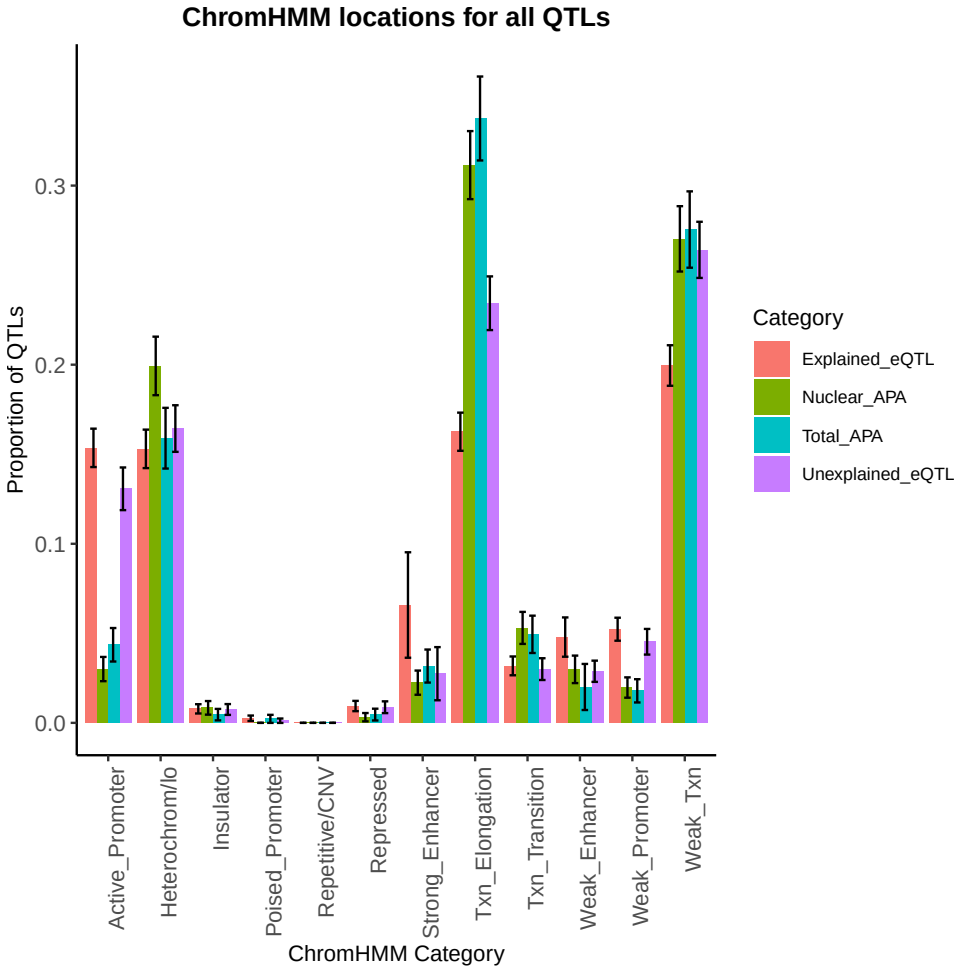


Figure 2.21: **Proportion of apaQTLs and eQTLs by Chromatin state** Bar plot showing the proportion of apaQTLs located in each of the 12 chromatin states from chromHMM. We find that the location profile of apaQTLs is more similar to that of unexplained eQTLs than that of explained eQTLs. Error bars represent the 95% confidence interval for each point estimate from bootstrapping 1,000 times.

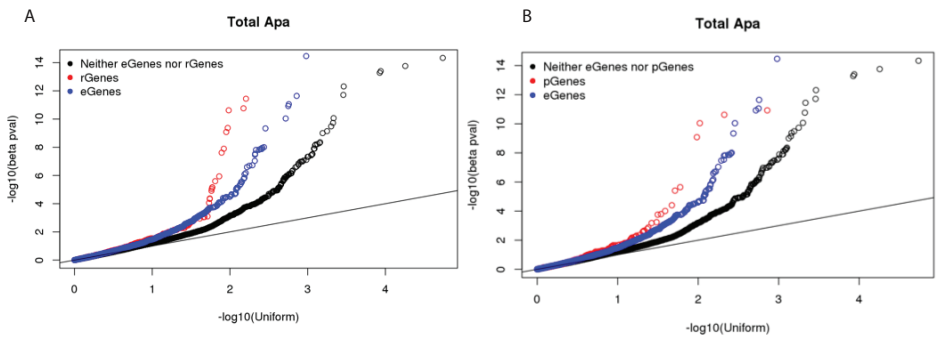


Figure 2.22: **Overlap between apaQTLs in total fraction and eQTLs, rQTLs and pQTLs supplement to Figure 2.4A (A)** QQ-plot showing the total apaQTL (adjusted) p-values separated by whether the corresponding gene has a ribosome occupancy QTL (red) or an eQTL (red). We see an enrichment for low apaQTL p-values in genes with either association. **(B)** QQ-plot showing the total apaQTL (adjusted) p-values separated by whether the corresponding gene has a protein expression QTL (red) or an eQTL (red). We see an enrichment for low apaQTL p-values in genes with either association.

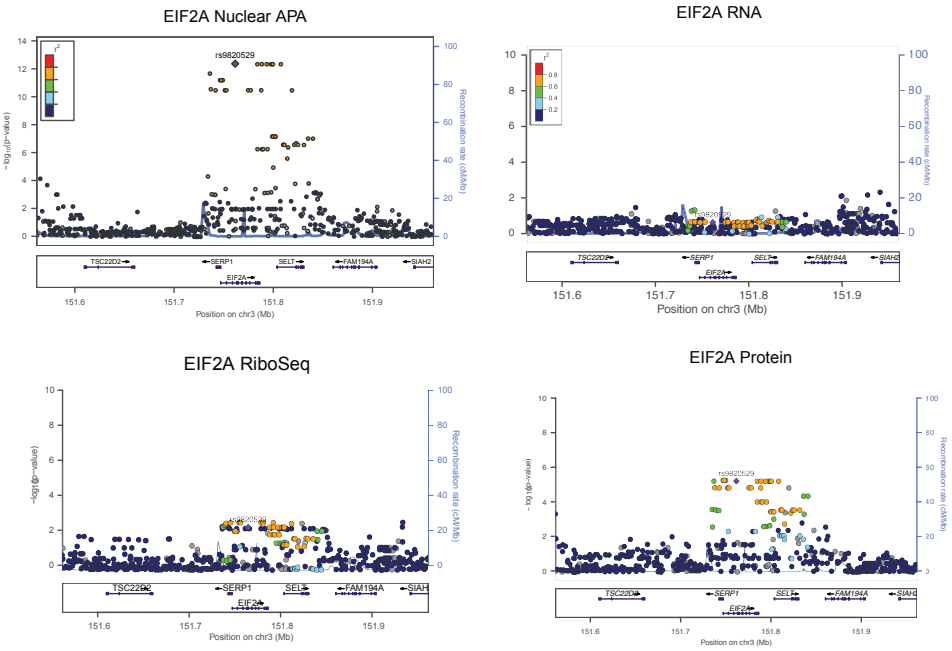


Figure 2.23: **LocusZoom plots for EIF2A molecular associations, Supplement to Figure 2.4B** LocusZoom plots for EIF2A apaQTL in Figure 2.4B along with associations with RNA expression, ribosome occupancy (ribo-seq), and protein expression as determined using normalized data from Li et al. 2016 [95]. LD patterns were colored according to the HapMap YRI lines.

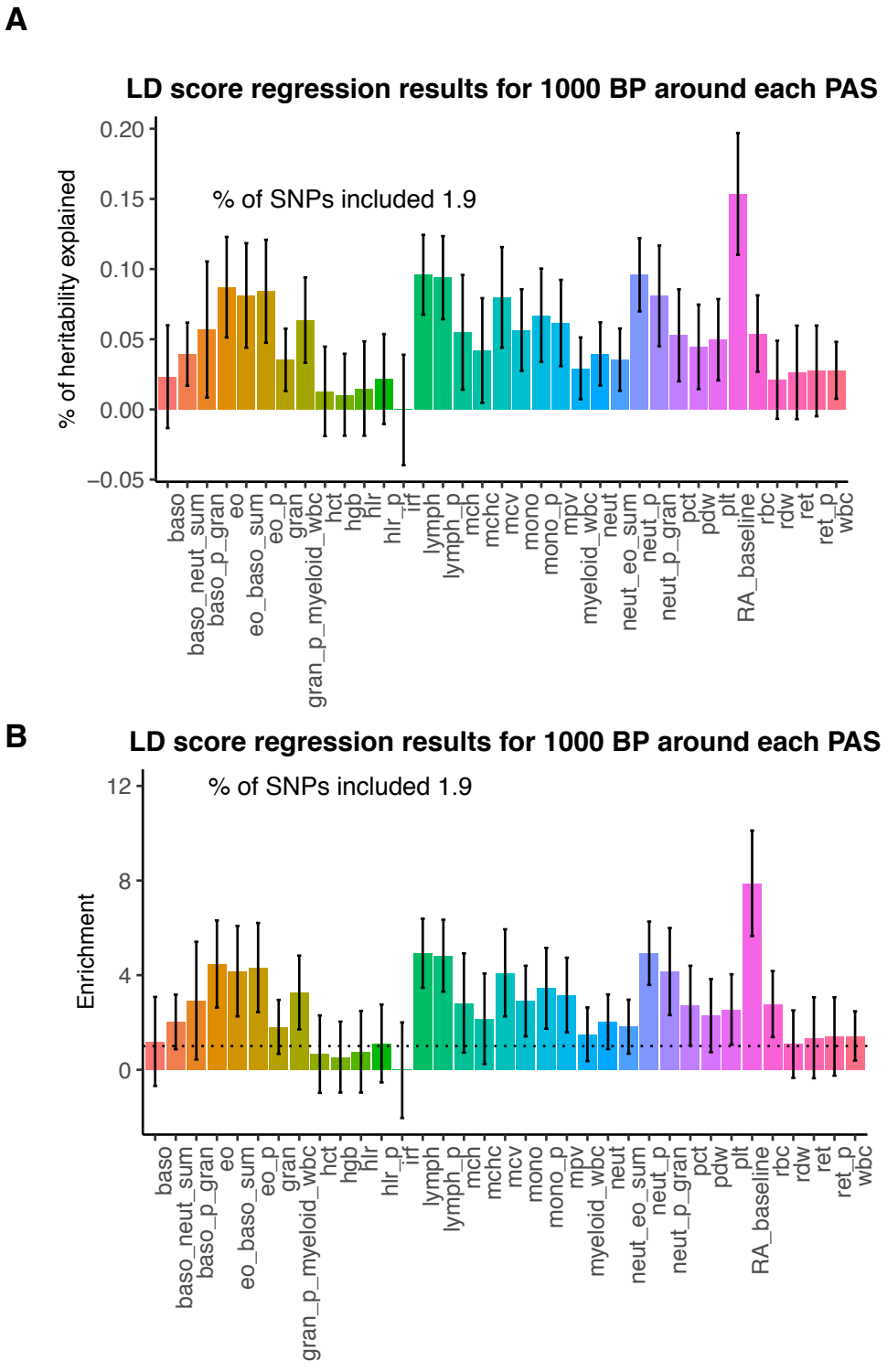


Figure 2.24: **LD Score regression enrichment estimates suggest that APA regulation is likely relevant for complex human phenotypes (A)** Percent of heritability explained by SNPs within 1kbp around each PAS.

Figure 2.24: (continued) Error bars represent +/- 1 standard deviation. Blood phenotype statistics published in Astle et el [5]. Rheumatoid arthritis statistics were obtained from Okada et al [129]. **(B)** Enrichment of heritability explained by SNPs within 1kbp around PAS for the phenotypes analyzed.

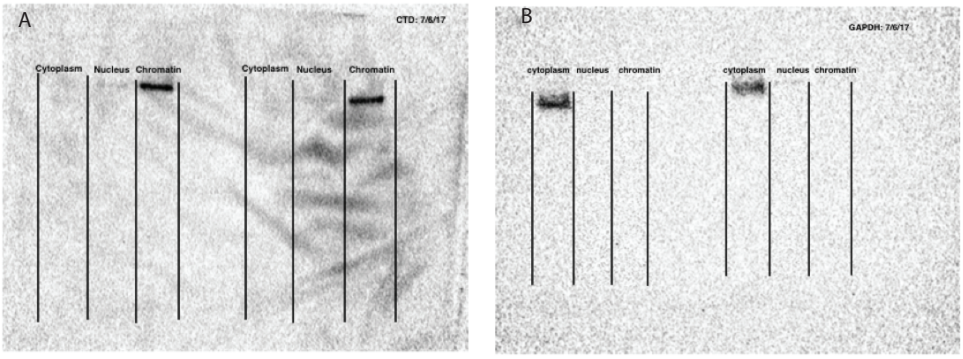


Figure 2.25: **Western Blots to demonstrate cell fractionation (A)** Western blot against Carboxyl terminal domain of RNA Polymerase II, photo captured at 10 second exposure. Blot is not used for quantification, but to validate cell fractionation. **(B)** Western blot against GAPDH to mark glycolysis in cytoplasm, photo captured at 25 second exposure time. Blot is not used for quantification, but to validate cell fractionation. Figure panels are modeled off Mayer and Churchman 2016, Figure 2 [111]

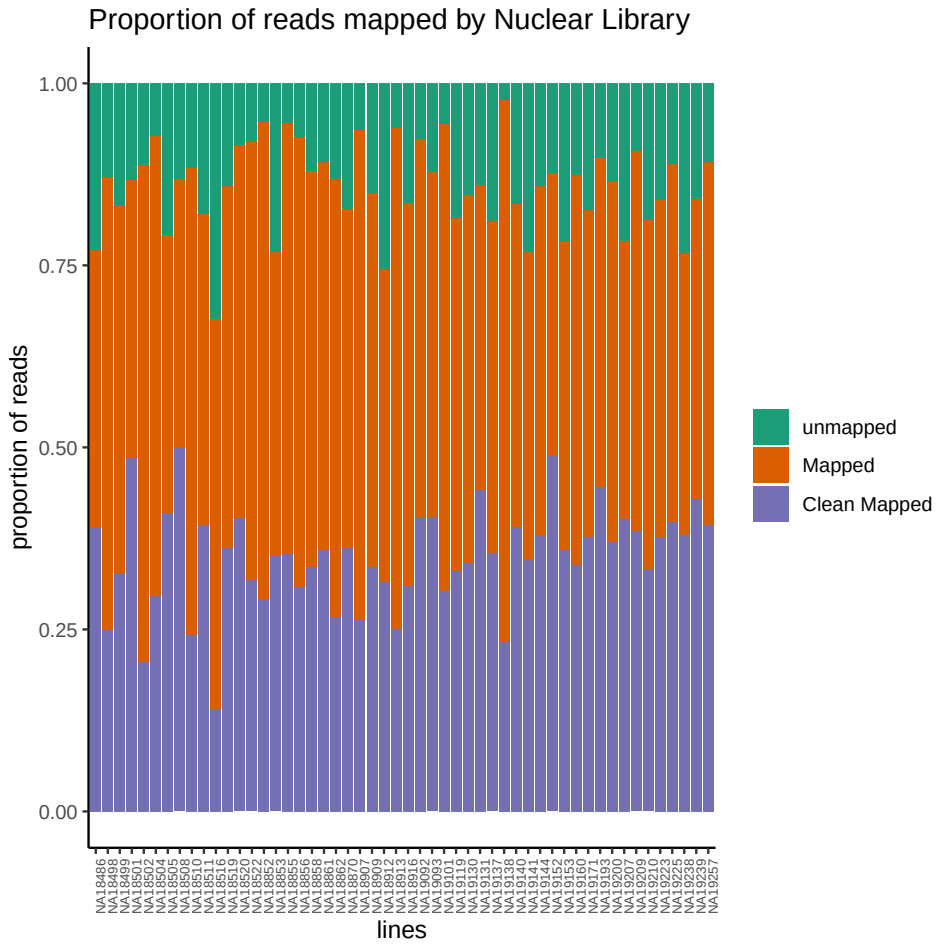


Figure 2.26: **3'-Seq read mapping proportions for the nuclear mRNA fraction** Proportion of reads that map to the genome (mapped) and the proportion of final reads used for analysis are cleanly mapped (Clean Mapped) by nuclear mRNA library. Cleanly mapped reads are reads that mapped successfully and passed the filtering for mispriming (MP) as described in the Methods.

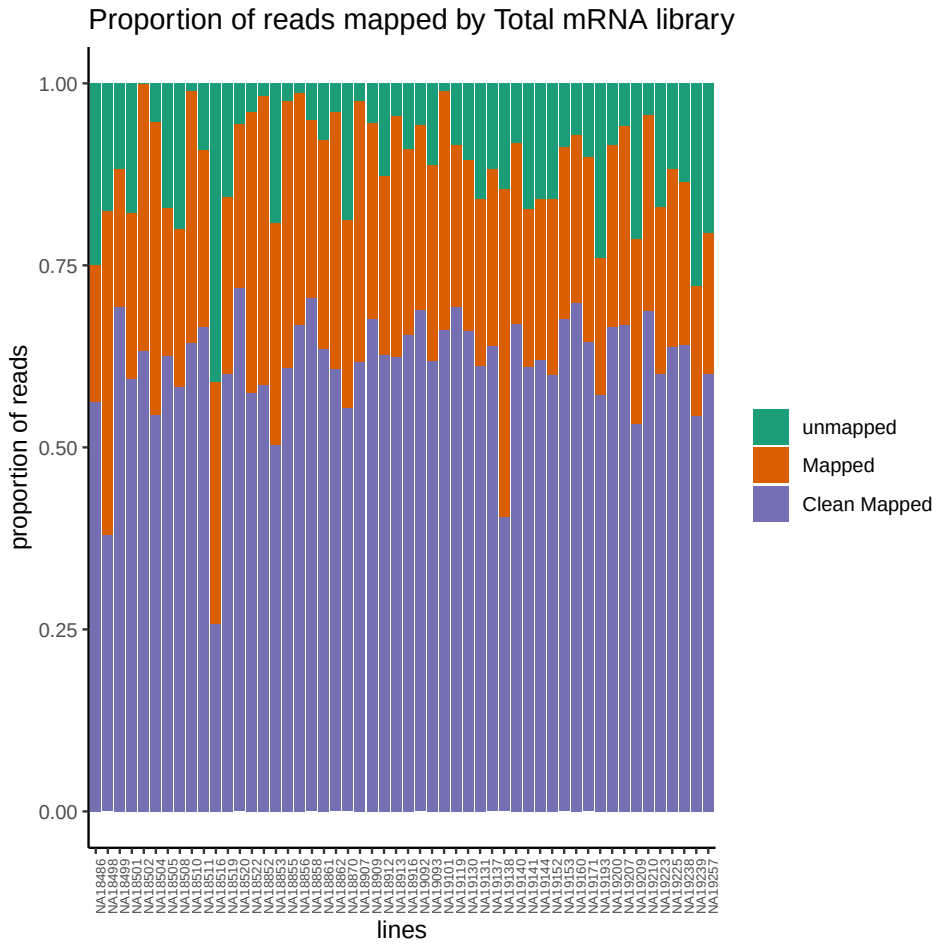


Figure 2.27: **3'-Seq read mapping proportions for the total mRNA fraction** Proportion of reads that map to the genome (mapped) and the proportion of final reads used for analysis that are cleanly mapped (Clean Mapped) by total mRNA library. Cleanly mapped reads are reads that mapped successfully and passed the filtering for mispriming (MP) as described in the Methods.

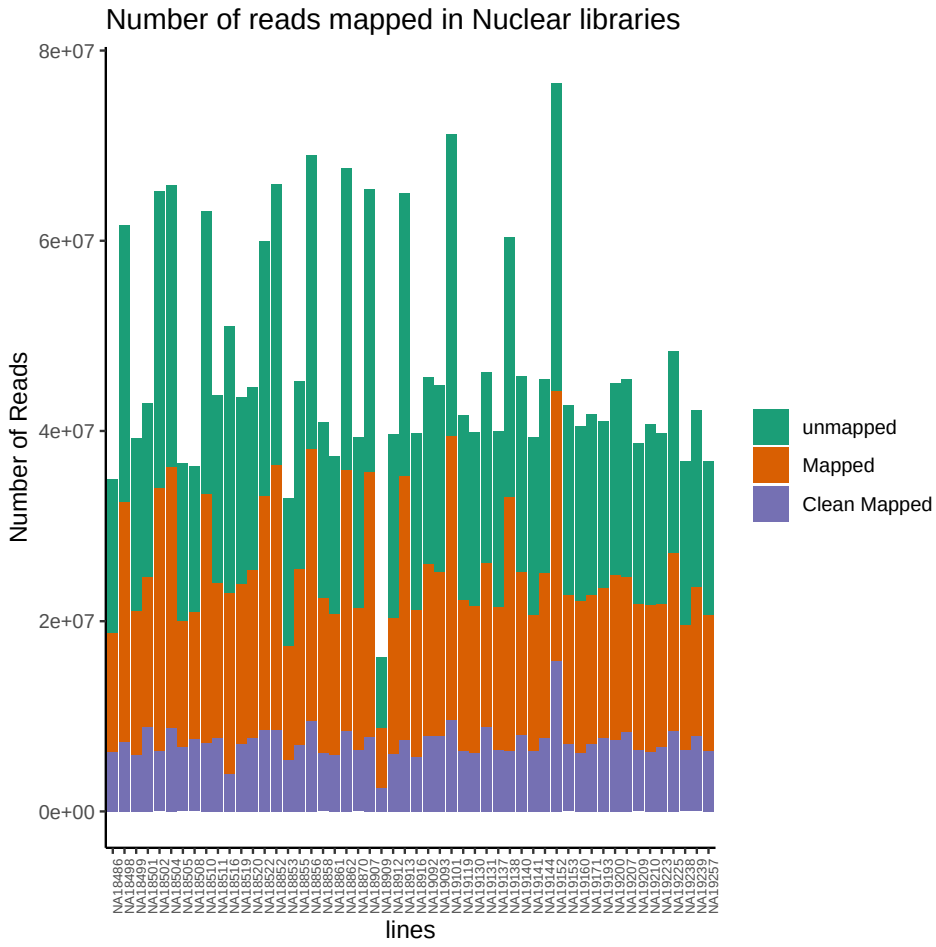


Figure 2.28: **3'-Seq reads mapping counts for the nuclear mRNA fraction** Total number of reads that map to the genome (mapped) and the number of final reads used for analysis that are cleanly mapped (Clean Mapped) by nuclear mRNA library. Cleanly mapped reads are reads that mapped successfully and passed the filtering for mispriming (MP) as described in the Methods.

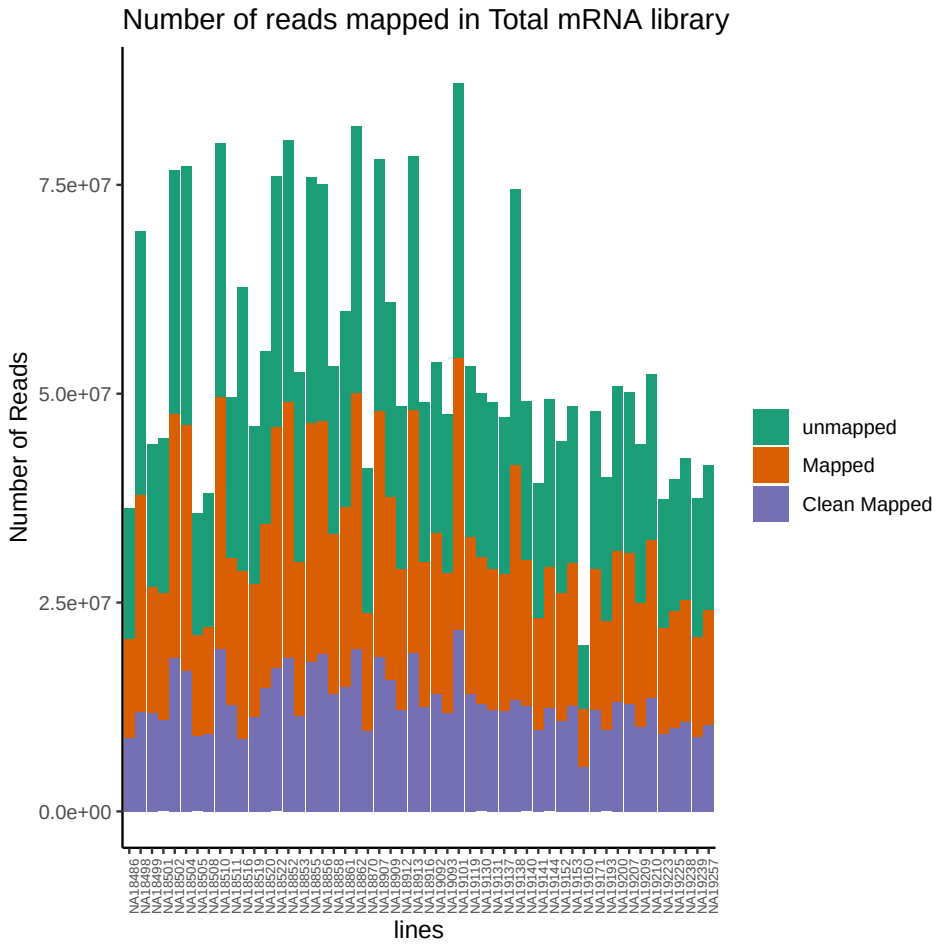


Figure 2.29: **3'-Seq reads mapping counts for the total mRNA fraction** Total number of reads that map to the genome (mapped) and the number of final reads used for analysis that are cleanly mapped (Clean Mapped) by total mRNA library. Cleanly mapped reads are reads that mapped successfully and passed the filtering for mispriming (MP) as described in the Methods.

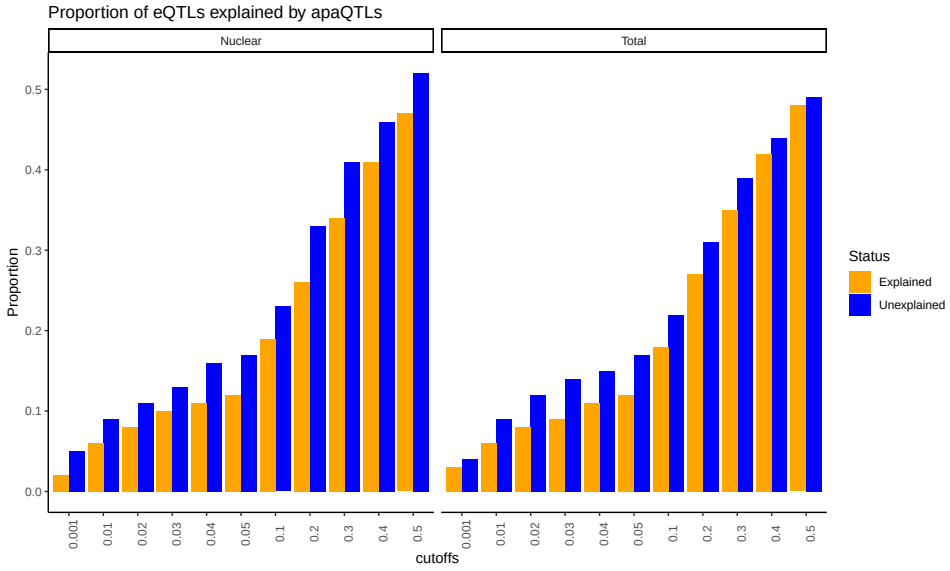


Figure 2.30: **Proportion of eQTLs explained by apaQTLs** Proportion of eQTLs putatively explained by apaQTLs separated by fraction. Expression QTLs could be explained by apaQTLs identified from both fractions. This observation is robust to apaQTL association p-value cutoffs. We observed that apaQTLs explain a slightly higher proportion of previously unexplained eQTLs. Explained/Unexplained status of each eQTL was determined previously in Li et al. 2016 [95]

2.9 Supplementary file 1

2.9.1 3' Sequencing of nuclear mRNA captures mRNA species independent of mRNA decay

To ensure that applying 3' Seq on the nuclear mRNA fraction would reflect polyadenylation usage of transcripts that have yet to be subject to decay, we verified that the nuclear mRNA 3' Seq captures features of nascent mRNA species prior to and independent from mRNA decay. To this end, we tested whether the ratio of nuclear to total mRNA 3' Seq reads correlates with measures of RNA decay. We reasoned that if nuclear mRNA captures mRNA species before they are subject to decay, then genes with more nuclear reads relative to total reads should have higher rates of mRNA decay. We used 4sU-seq (30m) data and RNA decay measurements collected in the same panel of lymphoblastoid cell lines (LCLs) as was used in this study as a proxy for mRNA rates of decay. The RNA decay and 4sU data were originally collected and processed in Pai et al. 2012 [136] and Li et al. 2016 [95], respectively. We further used RNA sequencing data collected in the same LCLs as used in this study and details regarding data processing can be found in Li et al. 2016 [95].

We computed a score reflecting the nascent transcription rate for each gene as the normalized 4sU count over the sum of the RNA-seq and 4sU counts. This is because 4sU captures nascent mRNA that were metabolically labelled with a modified uridine. After a fixed amount of time (30min in this case), the modified transcripts are sequenced. A positive correlation between 4sU/RNA and nuclear/total 3' Seq across genes suggests that the nuclear 3' Seq captures polyadenylation usage at an earlier stage of the mRNA lifecycle.

In Li et al 2016, the authors presented a relationship between the same nascent transcription rate and a measure relative mRNA decay rate. They reported a negative correlation between nascent transcription and relative decay, whereby genes with faster nascent transcription also show faster rates of decay. We show a similar relationship between decay rate

and our ratio of nuclear 3' Seq to nuclear and total mRNA 3' Seq, suggesting that we are capturing mRNA transcripts prior to mRNA decay in the nuclear fraction. To compute the correlations, we used the summary of the lm function in R.

Together, these correlations show that nuclear fraction 3' Seq captures information that is not captured in 3' Seq from the total mRNA fraction, and importantly, that the difference is biologically rather than technically driven. Thus, we were able to use 3' Seq data from both nuclear and total mRNA fraction to study how genetic effects regulate APA at multiple stages of the mRNA lifecycle. In particular, the observed difference between APA in nuclear versus total mRNA fraction supports the notion that if genetic effects were detectable only in the total mRNA fraction, we should suspect that the genetic effect drives variation in post-transcriptional regulation such as decay or export. This assumption is based on the premise that mRNA from the total fraction better reflect mRNA diversity subsequent to decay and export. Because we do not see many examples of genetic effects only identified in the total mRNA fraction, we propose that nearly all genetic effect drive variation in APA co-transcriptionally.

2.9.2 Intronic polyadenylation in other human tissues

In this study we used LCLs because of the rich molecular phenotyping that has been performed on the same cell lines. By collecting 3' Seq from cell nuclei we uncovered many more intronic PAS than expect. However, we are currently unable to validate whether these PAS are used in other human tissues because we are the first, to the best of our knowledge, to perform 3' Seq on mRNA from isolated nuclei in human cells.

That said, in order to estimate the extent to which intronic PAS we identified in the nuclear fraction are used in other human cell types, we turned to other APA studies that used a similar method to identify whole cell PAS. We reasoned that because total mRNA captures a small fraction of nuclear mRNA, it may be possible to use total mRNA to quantify

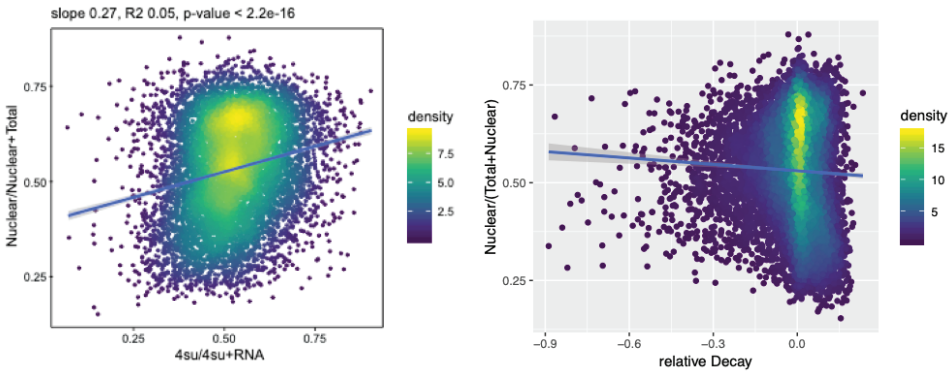


Figure 2.31: **Relationship between 3' Seq and nascent transcription (A)**Nuclear 3' Seq captures polyadenylation of nascent transcripts. The ratio of new mRNA to steady-state mRNA (x axis) are plotted against the ratio of 3' Seq reads from the nuclear fraction to 3' Seq reads from the total mRNA fraction (y axis). Slope, R^2, pvalue from a linear regression. **(B)** Nuclear 3' Seq captures polyadenylation of mRNA independent of mRNA decay. The relative decay rate of each gene (x axis) are plotted against the ratio of 3' Seq reads from the nuclear fraction to 3' Seq reads from the total mRNA fraction (y axis). Slope, R^2, pvalue from a linear regression.

the extent of intronic alternative polyadenylation in nuclei. For example, we found that 387 intronic PAS that were highly used in LCL nuclear mRNA were also detectable in LCL total mRNA. We can thus ask what fraction of these 387 intronic PAS also show evidence of usage in other cell-types from data collected by other studies on PAS. As baseline, we used 3' Seq usage data collected by Lianoglou et al., which include LCLs and four other cell-types (Breast, Ovary, Testes, Stem Cells). We found that about 10% of the 387 intronic PAS showed detectable usage in total 3' seq from LCLs collected by the Lianoglou study [96]. By contrast, around 5% of the intronic PAS showed usage in Breast, and Testes. Usage of 3' Seq data from another study performed by Derti and colleagues suggest that nearly 10% of the 387 PAS showed detectable usage [43]. Thus, these results suggest that there is at most a 2-fold difference in alternative polyadenylation in nuclei in other cell-types. While a 2-fold difference may appear large, we expect different cell-types to use different PAS depending on the specific genes that are expressed.

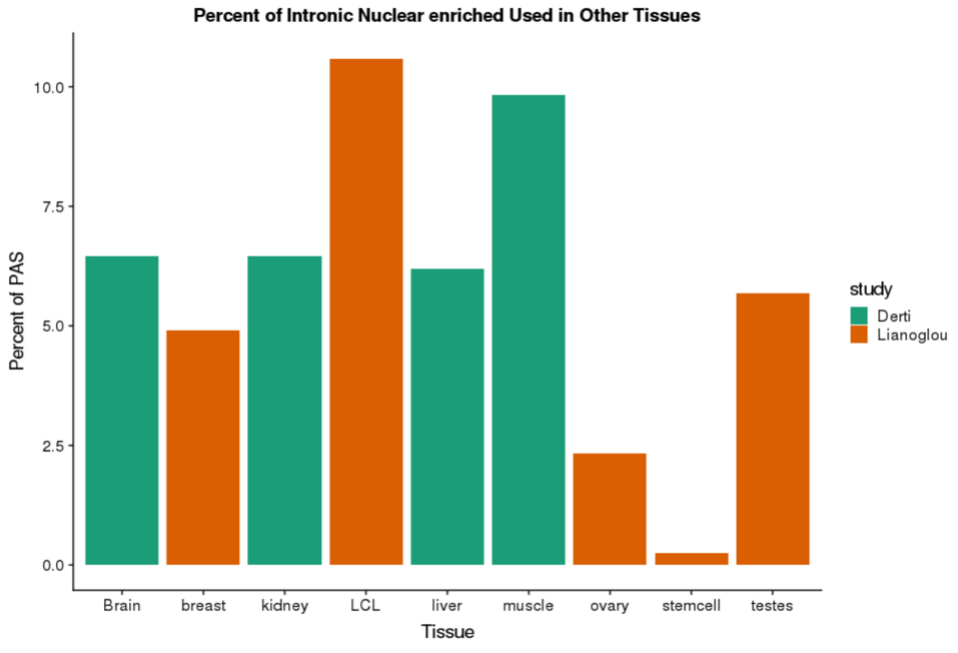


Figure 2.32: **Intronic PAS Discovered in other tissues** Intronic PAS enriched in the nuclear mRNA fraction of LCLs as detected in the total mRNA fraction of other human tissues. Barplot showing the percent of nuclear intronic PAS (of 387) discovered in whole cell 3' Seq from Derti et al. [43], or Lianoglou et al. [96] Bar for each tissue is colored by study in which the data was collected.

2.9.3 RNA binding motifs

3' UTRs are hotspots for RNA binding protein (RBP) motifs. When bound, RBPs can affect post transcriptional gene regulatory processes such as translation efficiency and nuclear export. We wanted to investigate whether genetic variants can impact APA by affecting binding of RBPs. To do this, we asked whether 3' UTRs with an apaQTL were more likely to be bound by an RBP than expected by chance. We downloaded eCLIP data for 25 RBPs collected by the ENCODE project in human K562 cells. We identified several RBPs enriched for genes with apaQTLs associated with 3' UTR PAS, but the overall enrichments were weak and are unlikely to explain the mechanism that underlie most apaQTLs. We did not see a similar enrichment for genes with intronic PAS apaQTLs. Interestingly, we found that the RNA binding proteins with the strongest enrichments are FUS and SAFB. These are intriguing result given the known function of FUS as a splice factor that guide nuclear export. We next asked if a genetic variant could be identified as an apaQTL due to differentially effects on one isoform but not the others. While we do not expect this to be the case genome wide, we do expect a small number of examples where a QTL could affect binding of an RBP and therefore isoform-specific post-transcriptional gene regulation. We identified 37 nuclear and 26 total apaQTLs overlapping eCLIP peaks. Of note, two apaQTLs disrupt binding for UPF1 which is a critical factor for nonsense mediated decay. A caveat to this analysis is the cell type specificity of RBP binding. eCLIP data is not available for LCLs.

2.9.4 Correlation between variance in ribosome occupancy and variance in APA

Variation in 3' UTR length can drive variation in translation efficiency. We wanted to test if this effect can be seen at the level of inter individual variation without requiring the existence of a QTL. We reasoned that if APA plays a role in modulating translation efficiency, then we would expect a correlation between APA variance and ribosome occupancy variance. When

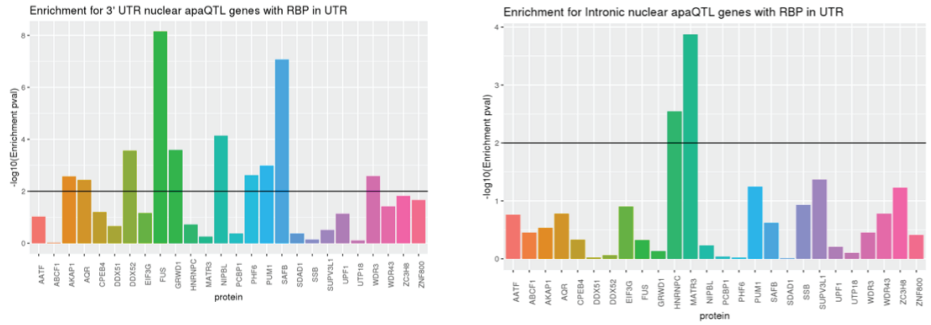


Figure 2.33: **Enrichment for RNA binding in K652 cells(A)**Enrichment for K562 cell RBP binding in 3' UTRs of genes with apaQTLs most strongly associated with a PAS in 3' UTRs compared to genes without apaQTL **(B)** Enrichment for K562 cell RBP binding in 3' UTRs of genes with apaQTLs most strongly associated with an intronic PAS compared to genes without apaQTL

we correlated the variance in usage for the most highly used PAS for each gene, we see a weak but significant positive correlation between APA variance and ribosome occupancy variance (Correlation $= 0.15$, $p < 2.2 \times 10^{-16}$).

2.9.5 Colocalization

In the main text we assert that APA can explain a proportion of the unexplained eQTLs, i.e. chromatin independent eQTLs. We primarily relied on correlation in order to draw this conclusion. However, to strengthen our claim, we used colocalization to ask if apaQTLs might generally be causal for the correlated eQTLs. To quantify the amount of colocalization between our apaQTLs and eQTLs, we used the COLOC package to test whether the apaQTL and eQTL associations share a causal SNP. The COLOC package estimates Bayes Factors for 4 alternative hypotheses. PP0: No association with either trait, PP1: No association with trait 1, PP2: No association with trait 2, PP3: Association with trait 1 and trait 2, two independent SNPs, and PP4: Association with trait 1 and trait 2, one shared SNP. If causal SNPs for an apaQTL and an eQTL is the same SNP, then PP4 is expected to be large ¿ 0.5. One limitation of COLOC is that it is very sensitive to sample size and tend to assign large posterior probability to PP0, PP1, PP2 when either of the QTL mapping

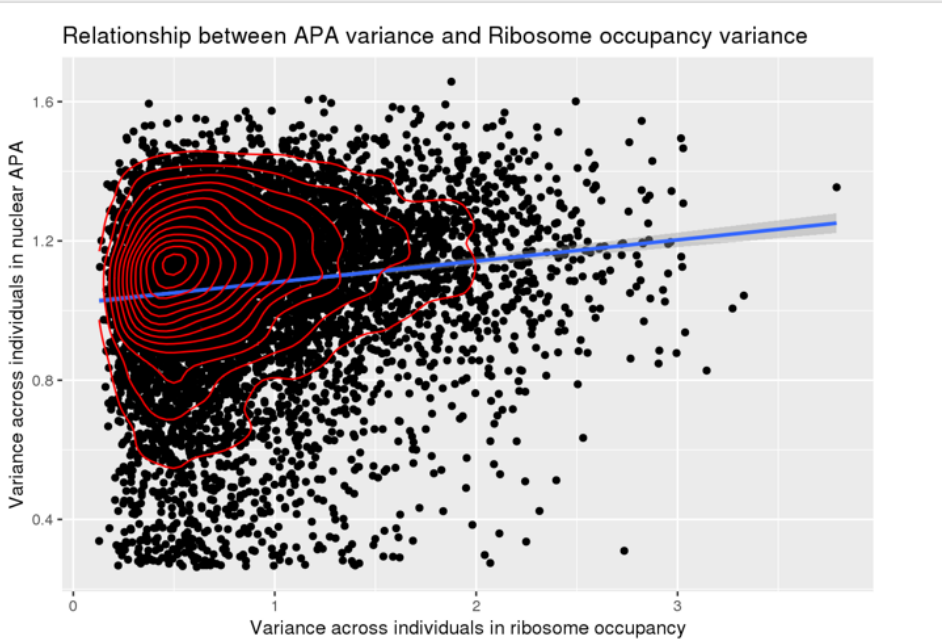


Figure 2.34: **Variance in APA and Ribosome Occupancy** Individual usage variance of the most highly used PAS in each gene (x axis) correlates with individual variance in ribosome occupancy (y axis) as measured in Li et al 2016.[95]

suffer from low power. This is because QTL mapping suffer from low power due to very small sample sizes compared to GWASs, for which coloc was designed for. To overcome this limitation, we used the ratio PP4/(PP3+PP4) to assess the colocalization probability instead of PP4/(PP0+PP1+PP2+PP3+PP4). To further increase power in our analysis, we used summary statistics from eQTLs identified on Geuvadis YRI LCL sample (n = 90) and used coloc to find colocalization between the eQTL signal and apaQTLs for the polyadenylation site (PAS) that is the most significant for the same gene. We expect this to be a lower bound for the actual number of colocalized eQTL-apaQTL SNPs because only one PAS for each gene is tested. Overall, we found that 33 genes had both and apaQTL and an eQTL and for which PP3+PP4 from coloc was 0.2 or greater. We found that the vast majority of genes (26, 78.8%) had a PP4/(PP3+PP4) value greater than 0.5, which indicates that the apaQTL and eQTL are more likely to share a causal SNP than not. Thus, we conclude that most apaQTLs that are determined to be eQTLs are likely to be causal, and further likely

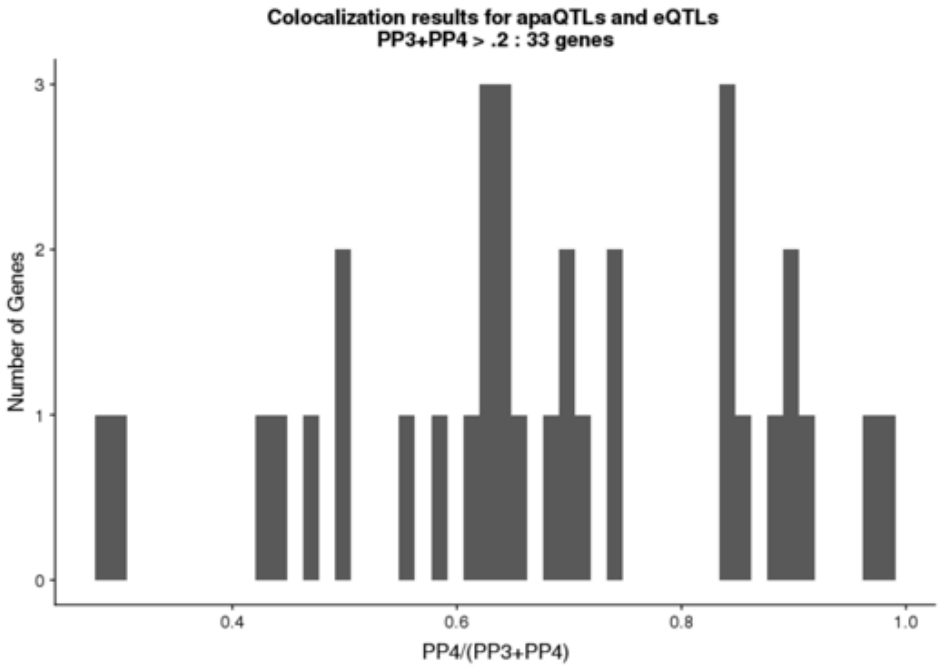


Figure 2.35: **Colocalization of apaQTLs and eQTLs** The apaQTL and eQTLs for the large majority of genes that have both are more likely to colocalize than not. Histogram of number of genes with an apaQTL and eQTL for different values of PP4/(PP3 +PP4).

explain all the SNP effect on gene expression.

2.9.6 Evaluating the robustness of our finding to false positives caused by mispriming

We took various measures to ensure that misprimed reads are not included in our analysis. For example, we include filters both at the read and PAS level according to previous reports using the same experimental protocol (methods). In order to test if mispriming could still be responsible for the PAS we identified, we have looked at the base composition around our PAS. The results are below with 10 base pairs up and downstream of the PAS (PAS are at position 10 on plot). We have separated PAS based on their location and on whether the PAS is annotated in polyADB. We found a very similar base pair composition for all PAS except for intronic PAS that are unannotated in polyA DB. This suggests there may

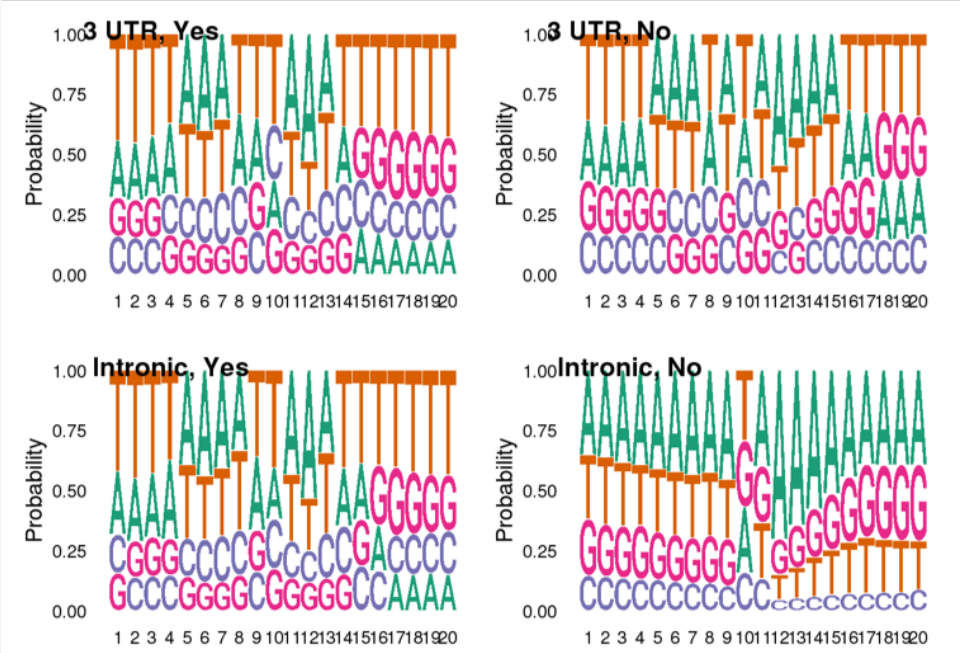


Figure 2.36: **Base Composition around PAS** Position weight matrices representing base composition 10 bps upstream and downstream of identified PAS separated by location and presence/absence of site in polyA DB.

be some amount of mispriming for intronic PAS that are not annotated in the polyADB. By quantifying the increase in A at nearby position around unannotated intronic PAS relative to annotated intronic PAS, we estimate that up to 20% of our unannotated intronic PAS may be explained by mispriming.

However, we believe that the vast majority of unannotated intronic PAS are likely to be real. To support this view, we found that of the 9,605 unannotated intronic PAS, 24.6% have a canonical polyadenylation signal site upstream of the PAS. This matched the fraction of intronic PAS that are annotated, and is significantly higher than background (which is about 0.24%). Furthermore, the location of the canonical polyadenylation signal site relative to the PAS location follows the expected distribution, which is 10-30bp upstream.

While we would argue that a 20% rate of mispriming is reasonably low, and removing more PAS would lead to many false negatives, we nevertheless decided to rerun our analysis after removing intronic PAS that have not been previously annotated, to make sure that

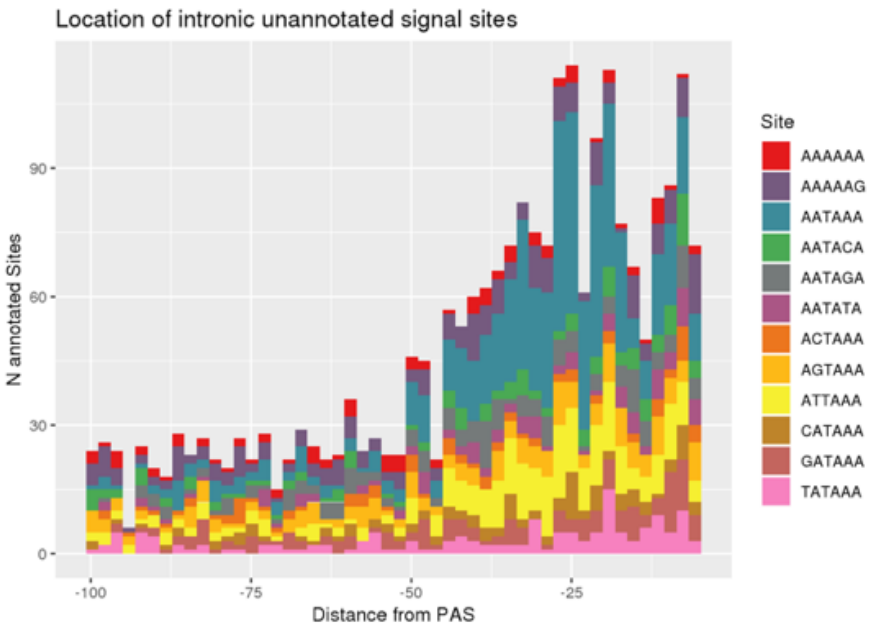


Figure 2.37: **Signal site distribution for intronic unannotated PAS** Stacked histogram of polyadenylation signal sites upstream of unannotated intronic PAS. Distribution similar in shape and structure to that in Figure 2.1D.

our results are robust to misprimed contaminates. We re-calculated the correlation between intronic effect sizes and eQTL effect sizes and found that the correlation is stronger than when the unannotated PAS are included (349 vs 357). This suggests that mispriming may be increasing noise.

We also found that the proportion of eQTLs that are significant apaQTLs does not change dramatically (18% vs 17.3% of unexplained eQTLs using the 0.05 cutoff).

Lastly, we found that nearly all apaQTLs that are not eQTLs but are associated with differences in translation and protein expression are not affected by the removal of unannotated intronic PAS (20 vs 25). Together these analyses suggest that even if our set of intronic PAS include some false positives, these PAS do not drive the main conclusions of our work.

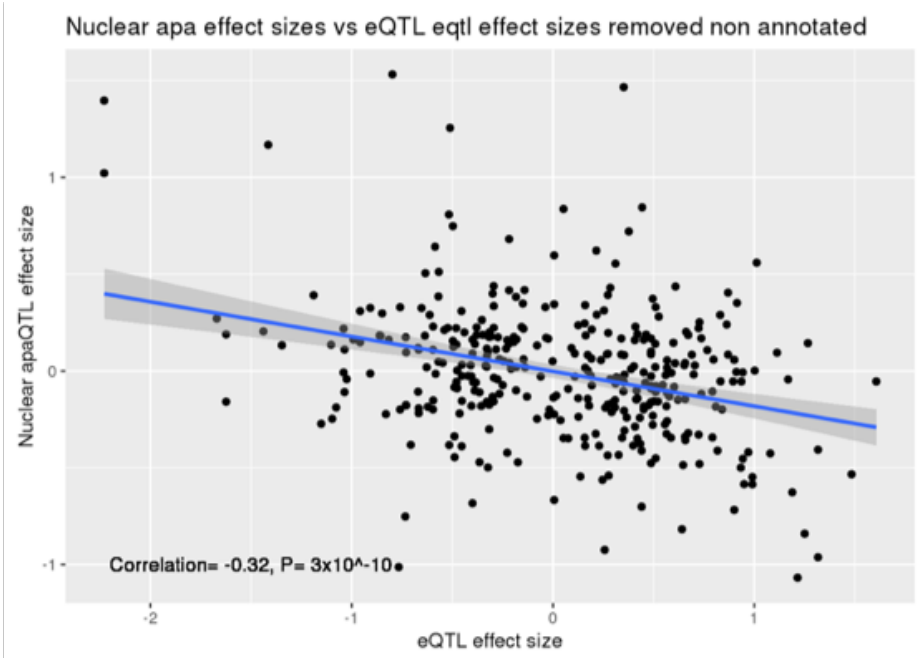


Figure 2.38: **Figure 2.3A without unannotated intronic PAS** Scatter plot of intronic apaQTL effect sizes after removing associations with unannotated intronic PAS plotted against their eQTL effect sizes. Supplemental to Figure 2.3A.

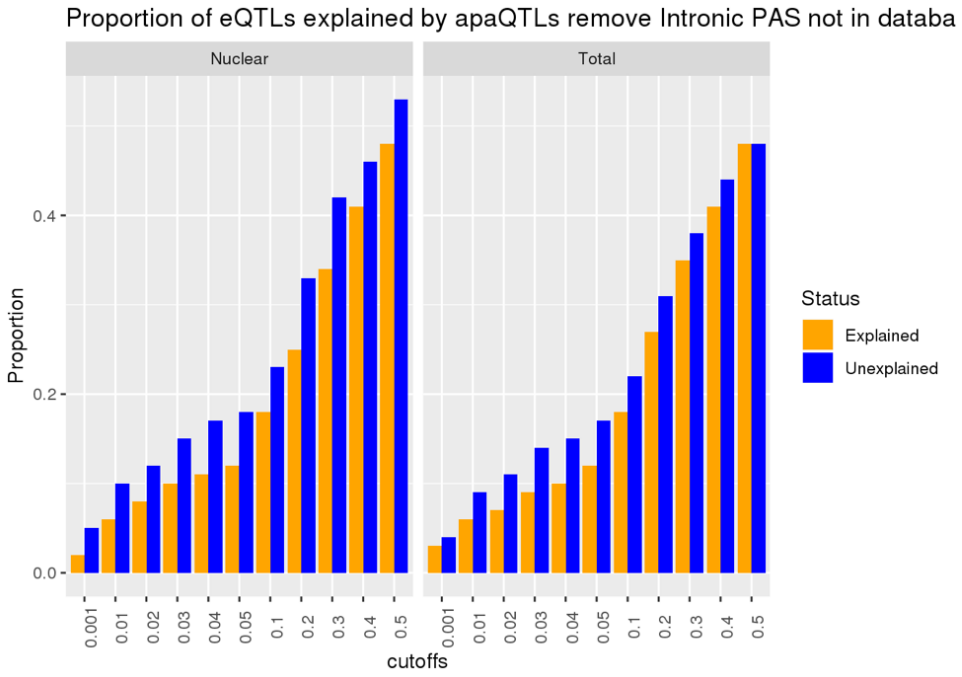


Figure 2.39: **Proportion eQTL explained without unannotated intronic PAS** Proportion of putatively explained by apaQTLs separated by fraction after removing associations with unannotated intronic PAS. Expression QTLs could be explained by apaQTLs identified from both fractions. This observation is robust to apaQTL association p-value cutoffs. We observed that apaQTLs explain a slightly higher proportion of previously unexplained eQTLs. Explained/Unexplained status of each eQTL was determined previously in Li et al. 2016.[95]

2.10 Supplementary Tables

Table 2.1: **Expression Independent eQTLs** (see supplementary file associated with this dissertation) apaQTL whose lead SNP is nominally associated with protein expression levels but not expression. Table includes p-value and slope for the associated between the lead SNP and nuclear APA usage, gene expression levels, protein expression levels, and ribosome occupancy (as measured using ribo-seq). mRNA, protein and translation data reported in Li et al. 2016[95].

Table 2.2: **Meta data** (see supplementary file associated with this dissertation) Library information for each Yoruba lymphoblastoid cell line, including sample, collection, and read information. Column names as described: Sample_ID: Sample ID, line: YRI Line,fraction: Molecular fraction, batch:3' Sequencing batch, fqlines: number of lines in fastQ file (used to calculate reads),reads: number of sequenced reads, mapped: number of mapped reads, Mapped_noMP:number of reads mapped after misprimmed reads are removed, prop_MappedwithoutMP, proportion of usable reads, Sex: Sex of YRI sample, Wake_Up: Date up cell line wakeup, Collection: Date of cell collection, count1:cell count measurements ($1x10^6$) , count2:cell count measurements ($1x10^6$), alive1: percent of cells alive calculated with trypan blue stain, alive2:percent of cells alive calculated with trypan blue stain, alive_avg: average of two percent alive measurements, undiluted_avg: average of two cell count measurements ($1x10^6$), Extraction: Date of mRNA extraction, Concentration: RNA concentration ng/ul, ratio260_280: RNA quality collected from nanodrop, to_use: amount of RNA input for 3' seq, h20: amount of water used for 3' seq, threeprime_start: data of library collection, Cq: quantification measurement from qPCR during 3' seq library preparation, cycles: cycles used for library prep, library_conc: concentration of 3' seq library (ng/ul).

CHAPTER 3

DIVERGENCE IN ALTERNATIVE POLYADENYLATION CONTRIBUTES TO GENE REGULATORY DIFFERENCES BETWEEN HUMANS AND CHIMPANZEES

3.1 Abstract[1]

Comparative functional genomic studies have shown that differences in gene expression between species can often be explained by corresponding inter-species differences in genetic and epigenetic regulatory mechanisms. In the quest to understand gene regulatory evolution in primates, the role of co-transcriptional regulatory mechanisms, such as alternative polyadenylation *(APA)*, have so far received little attention. To begin addressing this gap, we studied APA in lymphoblastoid cell lines *(LCLs)* from six humans and six chimpanzees, and comparatively estimated polyadenylation site *(PAS)* usage for 44,432 PAS in 9,518 genes. While APA is largely conserved in humans and chimpanzees, we identified 1,705 genes with significantly different PAS usage (FDR of 0.05) between the two species. We found that genes with divergent APA patterns are enriched among differentially expressed genes, as well as among genes that show differences in protein translation between species. In particular, differences in APA between humans and chimpanzees can explain a subset of observed inter-species protein expression differences that do not display corresponding expression differences at the transcript level. Finally, we focused on genes that have a dominant PAS, meaning the gene has a PAS that is used more often than all others for the same gene. Dominant PAS are highly conserved, and inter-species differences in dominant PAS are particularly enriched for genes that also show expression differences between the species. This study establishes APA as another key mechanism underlying the genetic regulation of

transcript and protein expression levels in primates.

3.2 Introduction

Humans and our close primate relatives exhibit a striking array of phenotypic diversity despite sharing homologous proteins with nearly identical amino acid sequences. Understanding how this diversity is propagated from genomic sequence to mRNA and protein necessitates an understanding of the regulatory mechanisms that occur before, during, and after transcription. Studying gene regulatory features in humans and other primates has long provided opportunities to understand genome evolution and function. For example, studies comparing patterns of epigenetic marks in primates have [8, 26, 117, 135]. Although many studies have focused on interspecies differences in the regulation of gene expression, fewer studies have addressed isoform-level variation, which contributes to differences in mRNA, translation, and protein levels between species.

The main mechanisms that contribute to mRNA isoform diversity are alternative splicing and alternative polyadenylation *(APA)*. Alternative splicing produces different combinations of coding sequences in mature mRNA and protein. APA occurs at genes that have more than one polyadenylation site *(PAS)* and can result in mRNAs with different coding sequences or variable 3' UTR lengths. Like alternative splicing, APA that occurs within the gene body can affect protein sequence and function [90, 140, 161, 179, 187, 201]. APA that occurs outside of the coding sequence, in the 3' UTRs, can lead to differential inclusion of protein-binding motifs that can affect translational efficiency, mRNA stability, and mRNA localization [115, 179]. Yet, despite its potential to produce tremendous variation in mRNA and protein regulation, few studies have explored the contribution of APA to regulatory divergence between species. Indeed, our current understanding of APA conservation in mammals

1. Citation for chapter: Mittleman BE, Pott S, Warland S, Barr K, Cuevas C, and Gilad Y. Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees. Manuscript in Prep

comes from few comparative studies of humans and rodents [4, 190]. However, these studies used sequence conservation rather than direct measurements of PAS usage to characterize APA [190]. Thus, it remains possible that many mammalian PAS are functionally divergent despite having similar sequences.

To gain insight into APA conservation between human and chimpanzees and understand how differences in APA contribute to gene regulation, we performed 3' sequencing (3' Seq) of mRNA isolated from nuclei collected from human and chimpanzee LCLs. We integrated PAS usage measurements with RNA-sequencing data collected from the same cell lines to understand the relationship between APA and gene expression. Finally, we used ribosome profiling and protein measurements previously collected in the same panel of human and chimpanzee LCLs to explore the effects of APA on protein levels [80, 191], reasoning that an understanding of how APA isoform usage varies among primates could help to explain why some human and chimpanzee genes are differentially expressed at the mRNA or protein levels, but not both.

3.3 Results

3.3.1 Describing alternative polyadenylation in human and chimpanzee LCLs

We performed 3' sequencing (3' Seq) of mRNA from 6 human and 6 chimpanzee lymphoblastoid cell lines *(LCLs)*, which we have previously used to study a variety of other functional genomic phenotypes [26, 80, 191, 207]. We collected mRNA separately from whole cells and isolated nuclei. The two cellular fractions serve as biological replicates, which we used to examine the quality of our data (Methods). In addition, by collecting data from isolated nuclei, we were able to capture polyadenylated transcripts before they became undetectable due to other regulatory processes, such as isoform-specific decay [118].

We mapped human 3' Seq reads to the GRCh38 reference genome and chimpanzee 3' Seq reads to the panTro6 reference genome [33, 163] (Methods). 3' Seq relies on a poly(dT) primer to target the poly(A) tail of mRNA molecules; however, it can also mis-prime by binding a sequence of genomic adenines. To account for mis-priming of off-target genomic sequences we removed reads that mapped to genomic regions containing $\geq 70\%$ adenine or 6 consecutive adenine bases in the 10 bp directly upstream of the mapped location [118, 164, 178] (Methods). In addition, we treated all ambiguous nucleotide positions as adenines to ensure that differences in reference genome quality did not bias the detection of polyadenylation sites *(PAS)* or mis-priming events (Methods). As expected, the filtered aligned sequences, merged by species, were enriched at transcription end sites *(TES)* and showed a similar distribution along orthologous 3' UTRs (Methods, Supplementary Figure 3.7). Next, we used a custom peak calling method to ascertain PAS in humans and chimpanzees separately (Methods) To compare PAS usage across species, we needed to identify the orthologous genomic regions of all PAS in our dataset, regardless of the species in which they were originally annotated. As we were unable to confidently identify orthologous PAS at base pair resolution (inferring synteny at base pair resolution in non-coding regions is challenging[21]), we extended each PAS by 100 bp upstream and downstream. We then used a reciprocal liftover pipeline to obtain an inclusive set of PAS regions with which we could confidently compare PAS usage between species (Methods).

To quantify PAS usage, we first assigned each PAS to a gene using the hg38 RefSeq annotation [151]. We then computed usage for each PAS in each individual as the fraction of reads mapping to one PAS over the total number of reads mapping to any PAS for the same gene (Supplementary Figure 3.8). We excluded PAS in lowly expressed genes or with less than 5% usage, as measurements from sparse data are highly susceptible to random error (Methods). We observed a strong correlation between PAS usage in mRNA from the nuclear and total cell fractions in all but one cell line (human NA18499; Supplementary

Figure 3.9). We re-identified PAS after removing all data from NA18499 and re-quantified PAS usage using nuclear 3' Seq data from 5 human and 6 chimpanzee LCLs. Using this analysis pipeline we identified a total of 44,432 PAS in 9,518 genes, which we used for all downstream analyses. On a genome-wide scale, we found that mean PAS usage is highly correlated between species (Pearson's Correlation, 0.9, $p < 2.2x10^{-16}$, Supplementary Figure 3.10). However, as expected, 41.8% of the variation in PAS usage (as explained by the top principal component of the data) is highly correlated with species (Pearson's correlation 0.99, $p = 2.95x10^{-8}$, Supplementary Figure 3.11).

We used a number of analyses to ensure our ability to detect PAS was not biased by gene expression level or species. If our ability to detect PAS was biased by gene expression, we would expect a positive correlation between gene expression level and the number of PAS we detected. In our data, the number of PAS per gene is negatively correlated with gene expression in both species (Supplementary Figure 3.12, Human: Pearson's correlation -0.17, $p < 2.2x10^{-16}$, Chimpanzee: Pearson's correlation -0.19, $p < 2.2x10^{-16}$). If our ability to detect PAS was biased by species, we would expect to identify more PAS per gene in one species over the other. This is neither the case genome-wide nor when we test each gene independently. We identified, on average, 3.87 PAS per gene in humans and 3.46 PAS per gene in chimpanzees. On average, per gene, the number of PAS in human minus the number of PAS in chimpanzee is 0.39 with a median value of 0 (Supplementary Figure 3.13. Moreover, as expected, the physical distribution of PAS across genes is conserved, with the majority of PAS located in 3' UTRs (17,688 (40.1%) in chimpanzee and 17,620 (40.0%) in human) and a considerable proportion located in introns (14,095 (31.9%) in chimpanzee and 14,119 (32.0%) in human) (Figure 3.1A).

To assess sequence conservation in PAS regions, we downloaded phyloP scores computed over 100 vertebrate genomes from the UCSC Genome Browser and calculated mean phyloP scores in PAS regions. Higher mean phyloP scores corresponds to regions of higher sequence

conservation, and thus, slower evolution [147]. Overall, sequence elements at PAS are more conserved than surrounding regions (Figure 3.1B, Wilcoxon rank sum test, $p < 2.2x10^{-16}$). This pattern holds independently for PAS in all genic locations other than introns (Supplementary Figure 3.14).

We identified 302 and 357 human- and chimpanzee-specific PAS, respectively (Methods). It has been previously shown that most PAS are directly preceded by one of 12 annotated sequence motifs that recruit cleavage and polyadenylation machinery to mRNA molecules as they are transcribed [12]. We asked if creation or disruption of a signal site motif could be responsible for species-specific PAS by mapping signal site motifs in both human and chimpanzee for each PAS region. Although human and chimpanzee PAS regions are equally likely to contain each of the 12 annotated signal sites (Figure 3.1C), only the top two most commonly used motifs, AATAAA and ATTAAA, are associated with increased PAS usage (Supplementary Figure 3.15). Thus, we considered only the presence or absence of these two motifs in subsequent analyses. Of the 302 human-specific PAS, 14 have human-specific signal sites and 6 have chimpanzee-specific signal sites. Of the 357 chimpanzee-specific PAS, 24 have a chimpanzee-specific signal site and 6 have a human-specific signal site. These numbers are small; still, species-specific signal sites are more abundant than expected by chance among species-specific PAS in human (5.7X, hypergeometric test, $p = 2.30x10^{-7}$) and in chimpanzee (8.3X, hypergeometric test, $p = 3.2x10^{-15}$), suggesting that signal site changes can explain a subset of differences in PAS usage. For example, we identified a chimpanzee-specific PAS about 1 kb upstream of a PAS used in both species in the 3' UTR of *MAN2B2*. The ancestral signal site conserved in chimpanzee is AATAAA; however, there has been a T to C transition in the human lineage [16] (3.16). This transition is likely responsible for the loss of PAS in humans.

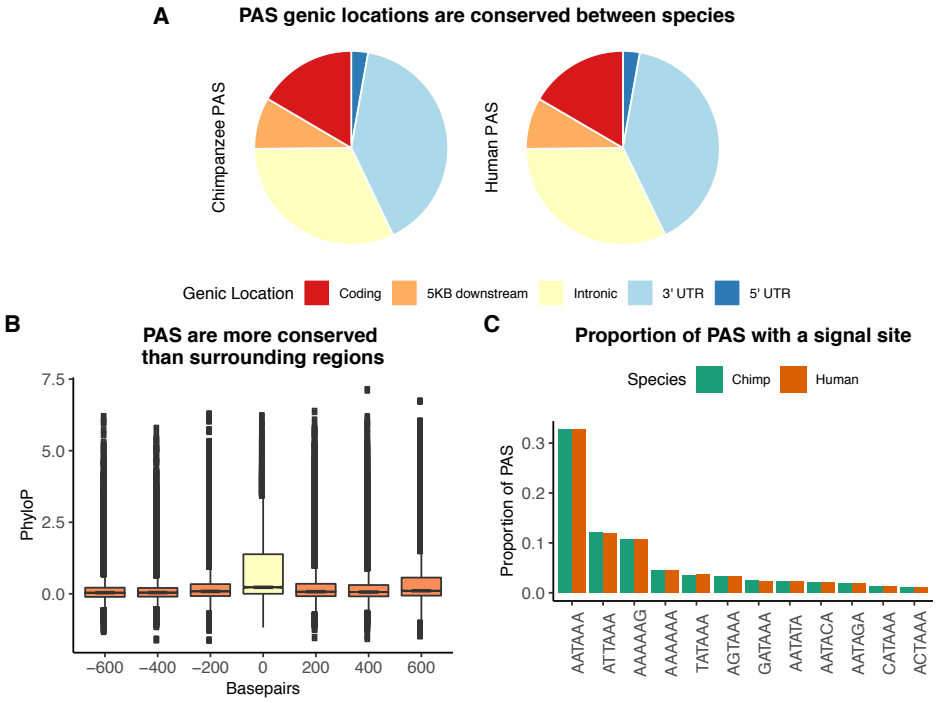


Figure 3.1: **Sequence conservation of PAS between humans and chimpanzees (A)** Genic locations for 44,074 PAS identified in chimpanzee (left) and 44,130 PAS identified in human (right). **(B)** Mean PhyloP scores for PAS regions (yellow) as well as three 200bp regions upstream and downstream (orange) **(C)** Proportion of human and chimpanzee PAS regions with each of the 12 annotated signal site motifs from Beudoing *et al.*[12]

3.3.2 Characterizing inter-species differences in PAS usage

While a few hundred PAS are species-specific, the majority of PAS (98.5%) were identified in both species. We thus sought to characterize quantitative differences in alternative polyadenylation (APA) patterns between human and chimpanzee by estimating the difference in usage of individual PAS in each species. To do so, we used the leafcutter differential splicing tool [94], which allowed us to tests for differences in normalized PAS usage fractions while accounting for gene structure (Methods). Using this approach, at an FDR of 5% we identified 2,342 PAS (in 1,705 genes) whose usage differs by 20% or more between the species (Figure 3.2A). We applied an arbitrary effect size cutoff to focus on larger inter-species differences, which are more likely to have functional consequences. The list of all PAS whose usage differs between the species, regardless of the effect size, is available in Supplemental Table 3.1.

To better understand the mechanisms that underlie inter-species differences in PAS usage, and the potential functional impact of such differences, we considered the APA data in different contexts. First, we noticed that the spatial distribution of differentially used PAS reflects the distribution of all PAS, namely differentially used PAS are most often located in 3' UTRs, followed by introns (Supplementary Figure 3.17). Within the 3' UTR, however, differentially used PAS are more frequently the first ones compared with PAS that are used similarly in the two species (Supplementary Figure 3.18, difference in proportion test, $p = 0.0015$). This pattern is intriguing, because changes in the usage of the first PAS in the 3' UTR may have the largest overall impact on the transcript length, and hence potentially the largest functional impact as well. However, it is also possible that we are more likely to detect differences in usage in the first PAS in the 3' UTR because this site is transcribed earlier, and our estimate of usage is relative to all other sites in each gene.

We therefore sought evidence that differences in PAS usage may have functional consequences. In a previous study, we identified genetic variants associated with variation in PAS

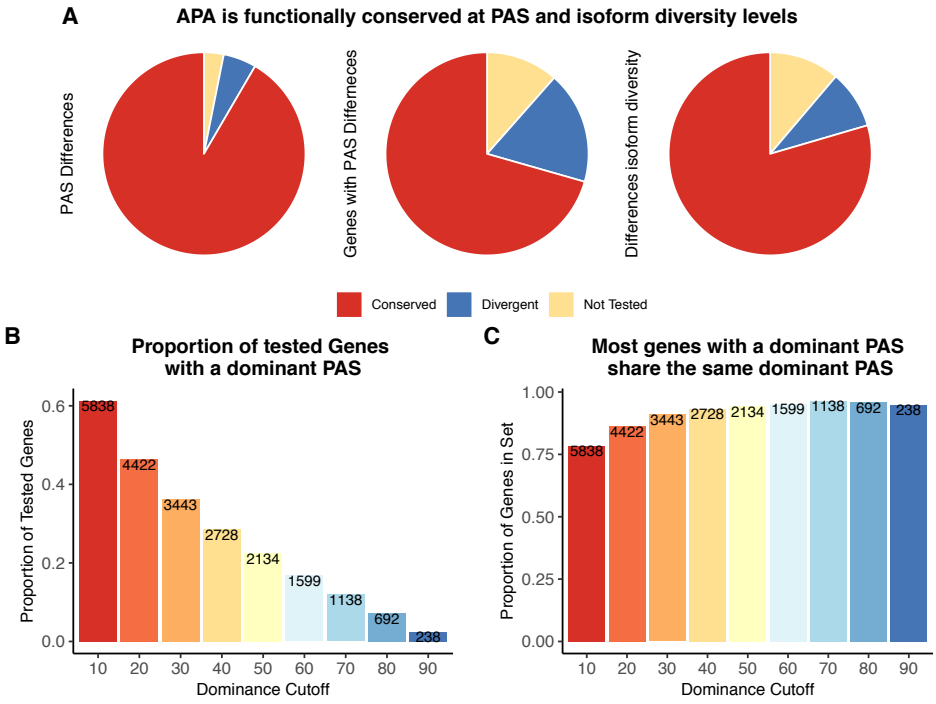


Figure 3.2: **APA is functionally conserved between humans and chimpanzees. (A)** Proportion of PAS and genes differentially used at PAS and isoform diversity level. (left) Divergent PAS are the 2,342 PAS differentially used at 5% FDR. Conserved are the PAS not differentially used at 5% FDR. Not tested PAS were removed from analysis by leafcutter tool. (middle) PAS level differentially used PAS reported at the gene level. Divergent genes are the 1,705 genes with PAS differentially used at 5% FDR. Conserved genes are the genes with no PAS differentially used at 5% FDR. (right) Divergent genes are the 881 genes with differences in isoform diversity between species at a 5% FDR. Conserved gene are genes without differences in isoform diversity. Genes with one PAS were not tested. **(B)** Proportion of tested genes with a dominant PAS in either species according to a range of cutoffs. Number of genes are reported in bars. **(C)** Proportion of the number of the number of genes with a dominant in either species that share the top used PAS according to each dominance cutoff. Number of genes with a dominant PAS is either species are reported in bars.

usage *(apaQTLs)* in a panel of 52 human LCLs [118]. We found that genes with inter-species differences in PAS usage are highly enriched for apaQTLs (160, empirical p-value based on 10,000 permutations $p = 0.001$, Supplementary Figure 3.19). This observation indicates that inter-species differences in APA usage can often be found in genes whose regulation varies also at the population level, generally suggesting relaxation of evolutionary constraint on the regulation of such genes. We next considered sequence divergence at PAS by obtaining phyloP scores for all PAS flanking regions (200 bp, as explained above). If many changes in PAS usage are genetically controlled, we expect genomic regions of differentially used PAS to be less conserved than regions flanking PAS sites that have similar usage. Indeed, differentially used sites are enriched for regions with negative mean phyloP scores (hypergeometric test, $p = 0.02$). This observation indicates that sequence divergence is often associated with differences in PAS usage, and that the majority of PAS usage in humans and chimpanzees may be generally conserved due to evolutionary constraint.

We next asked, more specifically, if signal site changes are likely to lead to differences in PAS usage. We addressed this question by performing two analyses. First, we focused on the 82 differentially used PAS with a signal site that is annotated in only one of the species. We found that the presence of a species-specific signal site is associated with increased PAS usage, as might be expected (human enrichment 3.82X, hypergeometric $p = 1.37x10^{-10}$, chimpanzee enrichment 3.02X, hypergeometric $p = 3.91x10^{-8}$). Second, we considered the presence of G/U-rich elements, which are known signals to the molecular machinery for polyadenylation [36]. Specifically, we considered the proportion of uracil bases in the PAS regions. Despite a high correlation in overall uracil content in both species (Pearson's correlation 0.99, $p < 2.2x10^{-16}$), the usage of PAS with greater uracil density in one species are more likely than expected by chance to be upregulated in that species (Chimpanzee 1.04X enrichment, hypergeometric test $p = 0.03$, Human 1.06X enrichment, hypergeometric test $p = 0.03$). Though species-specific signal sites explain a modest proportion of inter-species

differences in PAS usage, these cases demonstrate the link between sequence evolution and conservation of PAS usage between species.

3.3.3 The relationship between differences in alternative polyadenylation and gene expression

Our analysis to this point indicates that inter-species differences in PAS usage are often genetically controlled, but generally we have not found strong evidence that they are functionally important. We explored this further by considering the APA data in the context of gene expression data that we collected from the same 6 human and 6 chimpanzee LCLs (see Methods for data collection procedures and low-level analysis of the RNA-seq data). We found no meaningful correlation between inter-species differences in gene expression levels and changes in polyadenylation site usage *(ΔPAU)* in 7,462 genes for which we had both types of data (Pearson's correlation = -0.06, $p = 3.1x10^{-7}$, Figure 3.3A). We then separately considered the data for the 3' UTR and intronic PAS, because we previously found a different relationship between PAS usage in these genic regions and gene expression levels [118]. Indeed, we found that inter-species differences in the usage of intronic and 3' UTR PAS correlate with differences in expression effect between the species at an equal magnitude, but in opposite directions (Figure 3.3B). Increased usage of intronic sites is correlated with increased expression levels, while increased usage of 3' UTR sites is correlated with decreased expression.

We focused on 3,796 genes that were classified as differentially expressed between humans and chimpanzees at 5% FDR (Methods). We found that genes with at least one differentially used PAS between the species are more likely to be classified as differentially expressed than expected by chance (610 genes, 1.12X enrichment, hypergeometric test, $p = 3.18x10^{-5}$). Examining the subset of 610 genes, we observed a modest but significant negative correlation between differential expression effect size and ΔPAU when we considered all PAS (Pearson's

correlation= -0.15, $p = 0.0023$,Figure 3.3C). Separating the analysis by PAS genic location revealed, again, an opposite direction of the correlation between gene expression and the usage of either 3' UTR or intronic PAS (Figure 3.3D). These observations are consistent when we use PAS data based on 3' Seq data from whole cells instead of from the nuclear fractions, suggesting that the observed relationship is not due to nuclear export failure (Supplementary Figure 3.20, Methods).

To provide possible mechanistic insight into the relationship between PAS usage and gene expression, we identified AU-rich elements in 3' UTRs in both human and chimpanzee. AU-rich elements in 3' UTRs have been linked to destabilization of mRNA transcripts and translation repression [51, 121, 167]. We found that the 3' UTRs of genes that show an inter-species difference in 3' UTR PAS usage have a higher number (Wilcoxon test, $p < 10^{-16}$) and density (Wilcoxon test, $p < 10^{-16}$) of AU-rich elements compared with genes in which the 3' UTR PAS is similarly used in the two species.

3.3.4 Considering overall APA diversity

We explored the relationship between inter-species differences in APA and gene expression by using a different perspective. We hypothesized that we could gain more insight into regulatory variation by summarizing the PAS diversity for a given gene using a single statistic, rather than by analyzing the usage of each site separately. To do so, we measured isoform diversity using Simpson's D *(D)*, a metric traditionally employed by ecologists to measure taxon diversity between environments [122]. In our system, higher D values indicate that usage is spread more evenly across all PAS for a gene, while low D values suggest the one PAS is more dominant than others (Methods). As expected, in both humans and chimpanzees, D values are correlated with the number of PAS per gene (Supplementary Figure 3.21, 3.22, 3.23, human Pearson's correlation 0.62, $p < 2.2x10^{-16}$, chimpanzee Pearson's correlation 0.63, $p < 2.2x10^{-16}$).

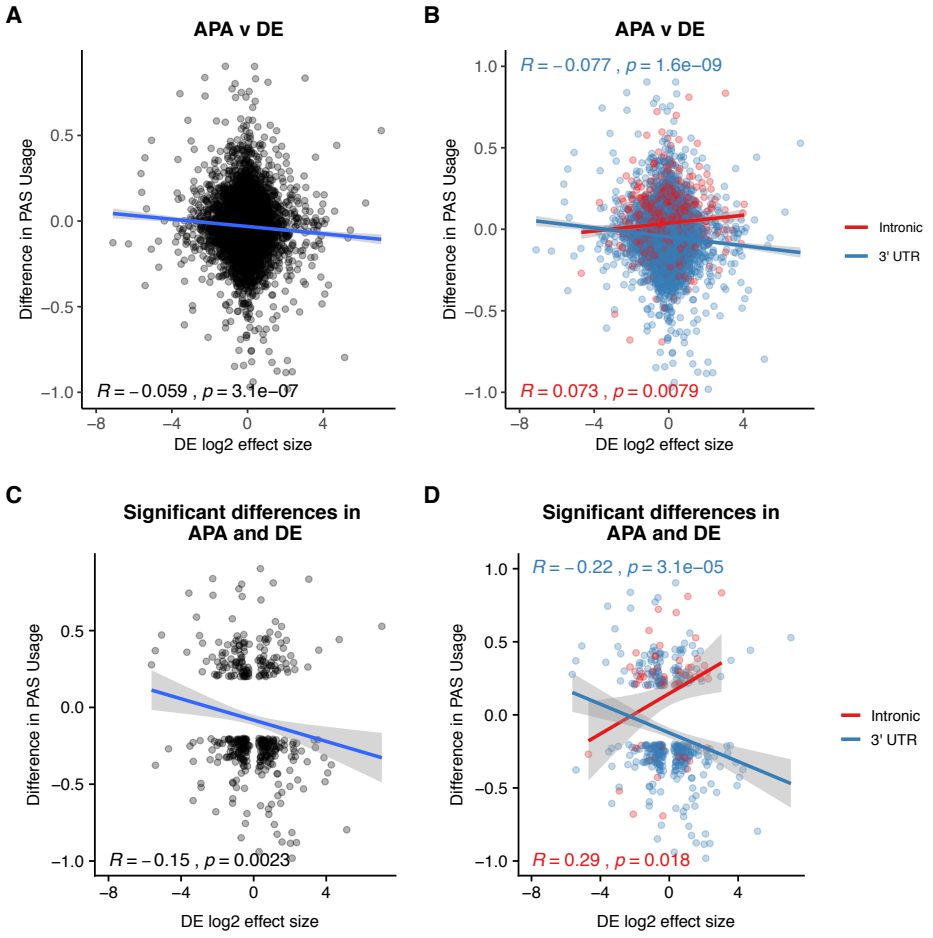


Figure 3.3: **PAS usage differences for intronic and 3' UTR PAS correlate with DE effect sizes at similar magnitudes but in opposite directions (A)** ΔPAU for top intronic or 3' UTR PAS per gene (Methods) plotted against differential effect size from differential expression analysis. **(B)** ΔPAU for top intronic or 3' UTR PAS per gene (Methods) plotted against differential effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. **(C)** ΔPAU for top intronic or 3' UTR PAS per gene (Methods) plotted against differential effect size from differential expression analysis. **(D)** ΔPAU for top intronic or 3' UTR PAS per gene (Methods) plotted against differential effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. In all panels, we calculated the linear regression and Pearson's correlation. In all panels, negative ΔPAU and DE effect sizes represent upregulation in chimpanzees. In panels B and D, we colored the points and regressions by genic location.

Using Simpson's D values calculated for each gene in each individual, we identified (at 5% FDR) 881 genes with significant differences in isoform diversity between species (Figure 3.2A; Methods). Of these, 426 are genes for which we did not previously detect an inter-species difference in PAS usage, indicating that Simpson's D is capturing an additional dimension of, or is more sensitive to, APA variation between species (Supplementary Figure 3.24, for example see Supplementary Figure 3.25).

We proceeded by focusing on genes with low isoform diversity, suggesting a single dominant PAS. We calculated a dominance metric for each gene as the difference in mean usage between the first and second most used PAS (we used different cutoffs to classify dominance; see Methods). We found that the classification dominant PAS is highly consistent across species; a result that is quite robust with respect to the approach used to classify PAS as dominant (Figure 3.2B,C). While the dominant PAS is the same for most genes in humans and chimpanzees, differences in usage of a dominant PAS are likely to contribute to differential APA with functional consequences between species more than differences in other PAS. Indeed, regardless of the specific cutoff we used to define dominant PAS, when the dominant PAS is not the same in humans and chimpanzees, the corresponding genes are more likely to be differentially expressed between the species compared with genes where the dominant PAS is the same in both species, (For cutoffs between .2 and .7, $p < 0.005$), and even compared with genes in which only a non-dominant PAS is differentially used ($p > 0.8$ for all cutoffs; Figure 3.4A,B).

In a previous study that collected mRNA from a larger panel of human, chimpanzee, and rhesus macaque LCLs, Khan *et al.* identified genes whose regulation likely evolves under directional selection in humans and chimpanzees [80]. We were able to consider RNA and protein expression data as well as APA data from 2,532 genes. We found that twenty-two of the genes with significant inter-species differences in APA at both the site level and in isoform diversity are among those whose regulation likely evolves under directional selection in the

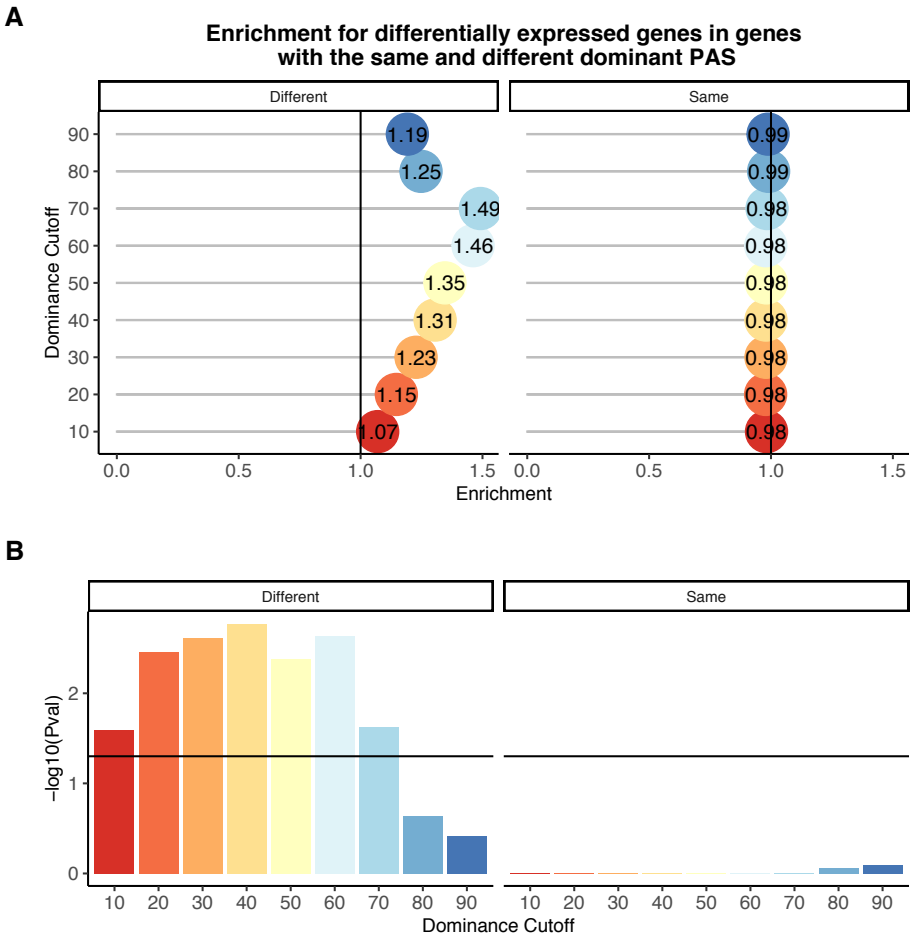


Figure 3.4: **Differences dominant PAS site between species likely drives differences in expression (A)** Enrichment of genes with the different (left) of same (right) dominant PAS by dominant cutoff in differentially expressed genes. **(B)** $-log_{10}(P-values)$ for enrichments in A calculated with hypergeometric tests. Horizontal line represents p-value of 0.05.

chimpanzee lineage, a 1.6X enrichment over what is expected by chance (hypergeometric test, $p = 0.015$). We did not find a similar enrichment when we considered genes whose regulation evolved under selection in humans, but the sample size is rather small.

3.3.5 Variation in APA and differences in protein expression

Given the well-characterized molecular connection between APA and the regulation of protein translation, we hypothesized that genes with inter-species differences in APA are also more likely to be differentially translated between the species [44, 51, 179]. To examine this, we obtained estimates of protein translation based on ribosome profiling data that were collected from human and chimpanzee LCLs by Wang *et al.* [191]. At a 5% FWER, Wang *et al.* identified 73 differentially translated genes between humans and chimpanzees. Genes with significant inter-species differences in isoform diversity are enriched among the differentially translated gene set ($p = 0.011$; Figure 3.5A,B).

We next investigated the relationship between ΔPAU in humans and chimpanzees and the effect sizes for differences in protein translation between the species [191]. Considering the most differentially used 3' UTR or intronic PAS per gene (Methods), we identified a significant correlation between inter-species differences in translation and ΔPAU for 3' UTR PAS, with a stronger correlation among genes with significant differences in both APA and translation (Supplementary Figure 3.26). As expected, and to some extent we view this as a control analysis, we did not identify a significant correlation between intronic PAS ΔPAU values and differences in translation (Supplementary Figure 3.26).

Given the apparent impact of PAS usage on protein translation, we next considered direct measurements of protein expression data from 3,391 genes in LCLs from humans and chimpanzees [80]. Using summary statistics from this study, we found 1,263 genes to be differentially expressed at the protein level between the species (FDR of 5%). As the protein measurements are restricted to these 3,391 genes, we do not have enough power to ask if

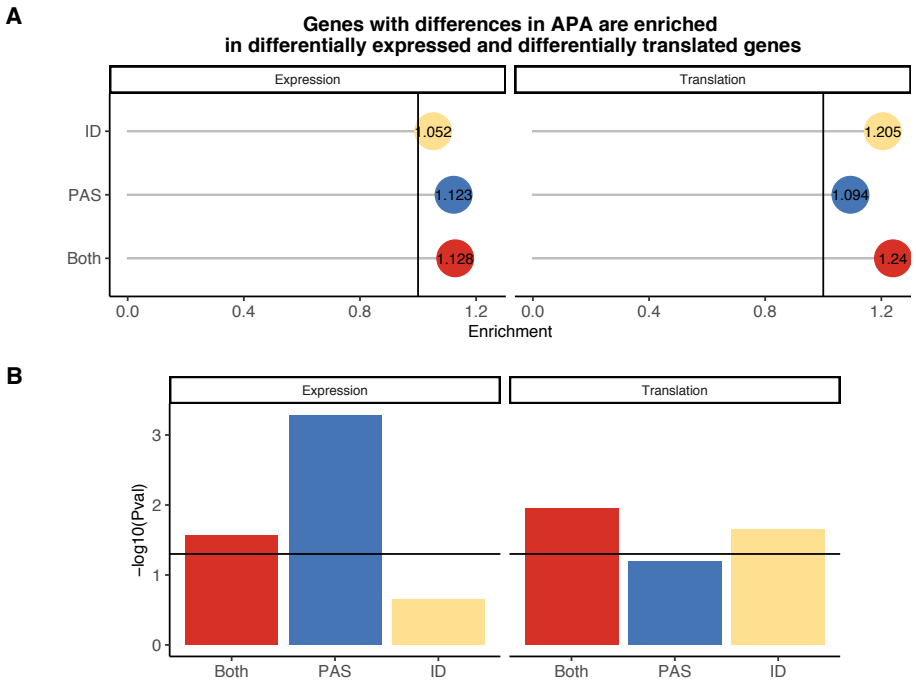


Figure 3.5: **PAS level differences in APA may drive differences in expression while isoform diversity differences likely drive translation differences (A)** Enrichment of genes with isoform diversity differences (ID), differences in APA at PAS level (PAS), or at both levels (Both) within differential expressed genes and differentially translated genes. Differentially translated genes reported by Wang *et al.*[191] **(B)** $-log_{10}(P - values)$ for enrichments in A calculated with hypergeometric tests. Horizontal line represents p-value of 0.05.

genes with inter-species differences in APA are also more likely to be differentially expressed at the protein level. However we did find a positive correlation between the absolute value of 3' UTR ΔPAU and the standardized number of ubiquitination sites for the same gene (Pearson's correlation, $R = 0.15$, $p = 5.0x10^{-7}$, Supplementary Figure 3.27, Methods), consistent with the observation that 3' UTR PAS are targets for the regulation of protein decay [47, 155].

Thus, we next focused on the 506 genes with significant inter-species differences in protein expression and an absence of corresponding difference of transcript expression levels that we also tested for differences in APA. Khan *et al.* reasonably hypothesized that inter-species differences in translation could account for the emergence of differences in protein expression levels when there are no regulatory differences at the RNA level, but they were unable to point to specific mechanisms. These genes are particularly interesting in the context of our current study, because APA which results in changes to 3' UTR length may be more likely to result in differences in protein expression without affecting the expression level of the mRNA.

Indeed, we found 76 genes with inter-species differences in APA that are also differentially expressed at the protein but not at the RNA level between humans and chimpanzees (Figure 3.6A,B). In these 76 genes, inter-species differences in PAS usage are enriched at the 3' UTR (Supplementary Figure 3.28). Finally, to assess whether APA contributes to differences in gene regulation by affecting translation efficiency or protein degradation, we asked whether genes with differential protein expression were also differentially translated. Of the 149 genes with significant differences in APA and protein expression, Wang *et al.* reported translation measurements for 142 [191]. Only 34 genes displayed significant differences in translation efficiency, suggesting that isoform-specific post-translational buffering is largely responsible for protein-level differences (Figure 3.6C,D).

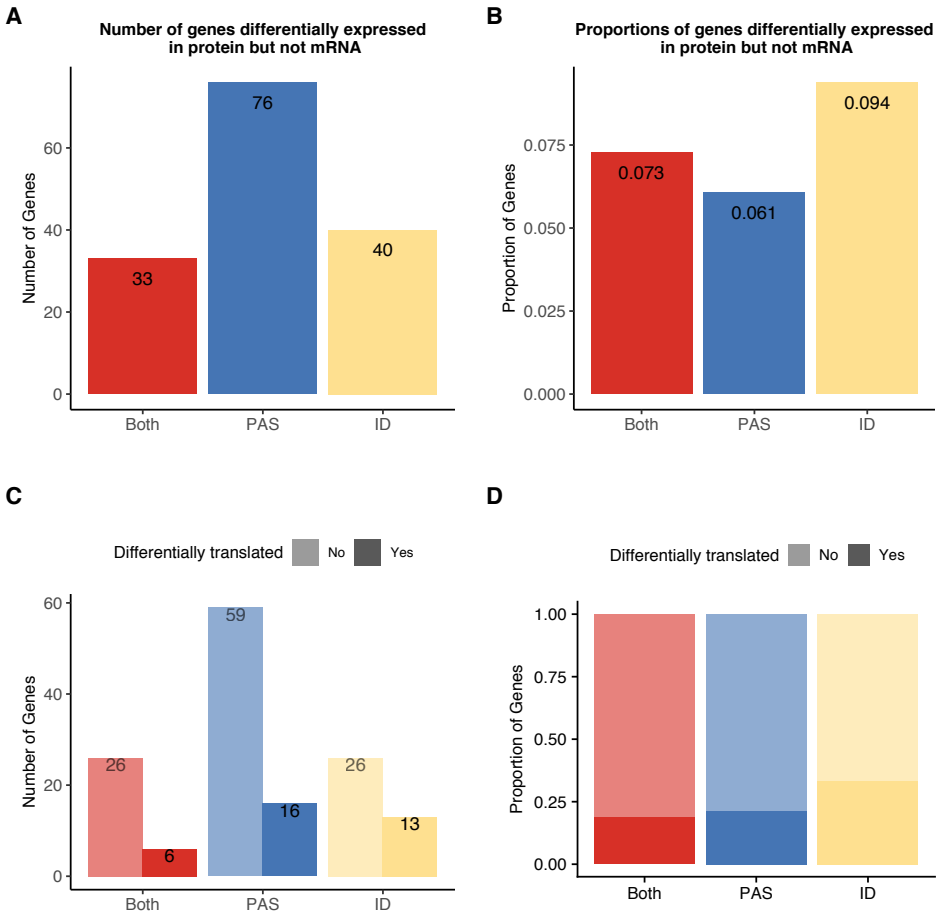


Figure 3.6: **APA differences explain genes differentially expressed at protein level but not in mRNA. APA likely mediates functional differences post translationally. (A)** Number of genes with isoform level differences (ID), differences in APA at PAS level (PAS), or at both levels (Both) differentially expressed in protein (5% FDR), but not mRNA (5% FDR). Genes with differentially expressed protein reported in Khan *et al.* [80] **(B)** Proportion of genes with differential usage at PAS level (1,251 genes), isoform diversity level (426 genes) or both (454 genes), differentially expressed in protein (5% FDR), but not mRNA (5% FDR). **(C)** Genes reported in A separated by genes differentially translated at 5% FDR. Differentially translated genes reported in Wang *et al.*[191] **(D)** Genes differentially expressed in protein but not in mRNA, colored by differences in APA. Proportion of genes in the set differentially translated at 5% FDR.

3.4 Discussion

Comparative primate functional genomics studies have contributed to our understanding of the gene regulatory processes that control genotype-phenotype relationships. The general framework starts with non-hypothesis driven research to characterize gene regulatory phenotypes in both human and non-human primates [17, 50]. In comparison to studying a regulatory phenotype in a human population, this approach is useful because, even at modest sample sizes, larger differences between primates make it easier to detect genes with biologically relevant differences [66].Once we understand the general properties and level of conservation of individual regulatory phenotypes, we can integrate data collected in the same cell lines or tissues to generate testable hypotheses for how variation in regulatory processes contribute to differences in physical traits. First, identification of genomic regions under adaptation in humans may point to causal mechanisms for human specific traits [58, 193]. Second, identification of heavily conserved genomic regions is critical because mutations in conserved loci likely have negative functional consequences and could be responsible for human disease risk [67]. With regard to this general framework, the results above represent the first step in understanding the evolutionary importance of gene regulation though APA. Greater characterization of APA variation in additional cell types will aid in the understanding of how variation percolates through the central dogma as well as how this variation contributes to differences in human traits.

We characterized alternative polyadenylation in humans and chimpanzees to begin to understand the role that co-transcriptional mechanisms play in the evolution of gene regulation in primates. We focused on the contribution of APA to interspecies differences in transcript and protein expression. Our group has studied a variety of gene regulatory phenotypes in primate LCLs [26, 80, 191, 207], and we, and others, have previously demonstrated that gene regulatory phenotypes in these cell lines recapitulate many regulatory patterns seen in primary tissues [28, 27, 79]. Not only did our use of primate LCLs allow us to circumvent

many of the practical and ethical issues associated with primate research, but it also allowed us to integrate 3' Seq and gene expression data from this study with previously collected ribosome occupancy and protein expression data from primate LCLs [80, 191].

APA is an important molecular mechanism with regard to both the evolution of gene regulation and physiological traits. On a long-term evolutionary scale, both 3' UTR length and the proportion of genes exhibiting APA have scaled with genome size and complexity. The expansion of APA has introduced biological complexity independent of an increase in the number of distinct genes [114, 115]. As usage of multiple isoforms have been maintained, it is likely that distinct isoforms have divergent functions acted on by balancing selection. For example, APA facilitates post-transcription regulation of a Drosophila Hox gene through maintenance of 2 isoforms differentially targeted by multiple miRNAs [142]. In terms of physiological traits, genome wide changes in APA during differentiation of stem cells to terminal cell types directs isoform specific gene regulation that is important for development in a range of species, including humans [64, 72, 93]. In addition, dysregulation of tumor suppressor genes through intronic polyadenylation is also known to contribute to cancer pathogenesis [46, 90]. We hypothesize that a further understanding of APA in primates will aid in the understanding of APA evolution and its contribution to human specific phenotypes.

We measured APA from 3' Seq data by calculating a ratio of isoforms terminating at one PAS compared to isoforms terminating at other PAS for the same gene. We then compared PAS usage ratios between species. To expand our understanding of APA conservation, we also calculated an isoform diversity statistic (Simpson's D) for each gene in each species. Because Simpson's D captures both the number of PAS isoforms and their usage, we were able to evaluate small regulatory changes spread across many PAS, rather than simply measure large changes at one PAS. While previous studies have used Shannon's Index to quantify isoform diversity [134, 191], we found Simpson's D to be less correlated with PAS number than Shannon's Index, making it less sensitive to the number of PAS per gene (Supplementary

Figure 3.22, 3.23). In addition, by placing more weight on dominant PAS, Simpson's D more closely mirrors our current biological understanding of APA, wherein dominant PAS play a larger role in downstream gene regulation [122].

In general, we found that both individual PAS usage and isoform diversity are highly conserved between human and chimpanzee. Consistent with comparative studies of APA in humans and rodents, which used genomic synteny to identify conserved PAS, we found higher conservation among genes with a single PAS [4, 190] and showed that sequence variation in PAS signal sites and the surrounding U-rich regions contributes to interspecies differences in APA [190]. Because we characterized APA in closely related primates, our study provides additional insight into APA divergence at both the gene level and the species level, revealing functional changes that contribute to differences in downstream gene regulation. For example, we observed that when genes use one PAS more often than others, the dominant PAS tends to be the same in both human and chimpanzee. It is likely that strong selection pressures have acted on these genes resulting in continual usage of the same dominant isoform. As a result, we can assume the dominant isoform is functionally important and alternative isoforms would be deleterious. The question remains, why are the non-dominant isoforms also conserved and at what level of expression do the alternative isoforms begin to impede gene function.

Out study also revealed that the majority of differentially used PAS between species are located in 3' UTRs. We showed that whereas increased intronic PAS usage is associated with increased mRNA expression levels, increased 3' UTR PAS usage is correlated with a decrease in mRNA expression. In a previous study, we found that human apaQTL alleles associated with increased intronic PAS usage were correlated with decreased mRNA expression levels [118]. We reasoned that transcripts terminating in introns were likely subject to nonstop decay, and by studying APA variation within humans, we were able to capture the effects of intronic termination. Because the current study characterized APA between humans

and chimpanzees, such effects were likely overshadowed by interspecies differences, which typically have much larger effect sizes [66]. Thus, usage of intronic PAS may contribute to down regulation of gene expression in some genes but in others, increased usage of intronic PAS result in truncated isoforms that do not contain 3' UTR cis regulatory elements that would signal mRNA decay. In these cases, the truncated isoforms are no longer targets of mRNA decay, causing certain genes to appear upregulated in RNA-Seq experiments.

As an alternative explanation for the results in this study, large effect sizes for differential usage of 3' UTR PAS could be driving the relationship between differential APA and differential expression. In line with this hypothesis, we pointed to increases in AU destabilizing elements and ubiquitination marks for genes with divergent 3' UTR PAS. Due the ratio nature of PAS usage, we may detect changes in the usage of intronic PAS solely as a mathematical consequence of changes in 3' UTR PAS. Functional follow-up on the genes with PAS detected as differentially used between and within species would be necessary to explore the relative importance of each of these regulatory pathways and to disentangle the results from both studies.

Indeed, APA is not the first molecular phenotype wherein a within-species study revealed alternative regulatory models compared to an interspecies analysis. For example, Pai *et al.* reported tissue-specific differential methylation to be almost exclusively inversely correlated with gene expression patterns between human and chimpanzee [48, 135, 194], whereas Banovich *et al.* discovered genetic variation associated with methylation variation *(meQTLs)* that was both directly and inversely correlated with eQTLs.

In several previous studies from our group and others, the authors estimated the proportion of variation in expression explained by the regulatory mechanism of interest. Specifically, they tested for differences in expression before and after accounting for another regulatory mechanism with formal mediation analyses [15, 19, 26, 49]. Similar mediation analyses may have helped us to better understand how usage of intronic and 3' UTR PAS contribute to

differences in mRNA expression. However, in our study, APA was measured using ratios of alternative mRNA isoforms. Thus, effect sizes for APA and differential expression are on different scales and we cannot use a mediation approach to formally calculate the proportion of expression variation explained by APA. However, we are generally convinced that APA contributes to differences in mRNA expression overall because 764 of 3,796 (20.1%) differentially expressed genes also have significant differences in APA.

Even though genes are ultimately expressed as proteins, many studies measure mRNA expression as a proxy for gene expression. As a justification for this approach, authors point to the fact that after accounting for technical considerations, the correlation between mRNA levels and abundance in quite high genome wide [22, 38]. However, we also know that on the level of one gene across individuals or tissues, the correlation between mRNA and protein measurements are much lower [11, 22]. This suggests that a number of molecular mechanisms decouple mRNA and protein levels post transcriptionally. As evident by the fact that we are still unable to predict protein levels from mRNA, we do not fully understand the post transcriptional mechanisms that shape the proteome [22].

Within human populations and between primates there are large number of genes differentially expressed in mRNA, but not in protein. Directly measuring translation levels for these genes has demonstrated that post-translation protein buffering can explain the decreased variation at the protein level [11, 191]. Conversely there are also genes that are more variable in protein than in mRNA [11, 32, 80]. Our previous work demonstrated that some protein specific QTLs are also highly correlated with differences in APA [118]. Here, we expanded this analysis and demonstrated that genes differentially expressed in protein, but not in mRNA between human and chimpanzee, have divergent APA patterns. We also concluded that the divergent protein levels are likely due to post-translational molecular mechanisms. While we cannot directly test the mechanism here, by differentially including RNA and protein binding motifs, APA could leads to variation in protein levels as a consequence of

protein auto-regulation [22, 40, 125]. Alternatively, APA could contribute to temporal and spacial differences for proteins, that would modify our ability to quantify protein with traditional techniques [22, 179]. In conclusion, a better understanding of co-transcriptional gene regulatory mechanisms, such as APA, may point to additional mechanisms contributing to the decoupling of mRNA and protein abundance.

3.5 Methods

3.5.1 Cell culture and collections

We grew 6 human and 6 chimpanzee Epstein-Bar virus transformed lymphoblastoid cell lines *(LCLs)* in glutamine depleted RPMI [RPMI 1640 1X from Corning (15-040-CM)], completed with 20% FBS, 2mM GlutaMax [Gibco (35050-061)], 100 IU/ml Penicillin, and 100 ug/mL Streptomycin. We cultured all cells at 37C at 5% CO2. We passaged each cell line a minimum of 3 times then maintained cells at $1x10^6$ cell per mL in preparation for collection. Cell line numbers and details can be found in Table 3.2. The human lines were derived from Yoruba individuals collected as part of the HapMap project and can be ordered through the Coriell Institute [70]. Chimpanzee LCLs were originally transformed from individuals from the New Iberia Research Center (University of Louisiana at Lafayette), Coriell IPBR repository and Arizona State University [80]. The cell lines have previously been used for similar studies of primate gene regulation [26, 80, 191, 207].

Once all cells lines reached $1x10^6$ cell per mL, we used the collection and RNA extraction method detailed in Mittleman *et al.* to extract whole cell and nuclear mRNA. Briefly, we collected 30 million cells in two 15 million cell aliquots. We extracted nuclei from one aliquot per line using the nuclear isolation protocol outlined by Mayer and Churchman [111]. We extracted mRNA in two fraction- and species-matched batches, using the miRNeasy kit (Qiagen) according to manufacture instructions, including the DNase step to remove

genomic DNA. We quantified mRNA and tested quality using a nanodrop. Details of mRNA processing for each line, including concentrations and quality can be found in Table 3.2.

3.5.2 3' Sequencing to identify PAS and quantify site usage

We generated 3' sequencing *(3' Seq)* libraries from whole cell and nuclear-isolated mRNA from 6 chimpanzee and 6 human individuals using the QuantSeq Rev 3' mRNA-Seq Library Prep Kit [120] according to the manufacturer's instructions. We sequenced all libraries on the Illumina NextSeq500 at the University of Chicago Genomics Core facility using single-end 50 bp sequencing.

We mapped human 3' Seq libraries to GRCh38 [163] and chimpanzee libraries to panTro6 [33] using the STAR RNA-seq aligner with default settings [45]. Similar to our previous work, we removed reads with evidence of internal priming resulting from the poly(dT) primer. We filtered reads proceeded by 6 As or 7 of 10 As in the base pairs directly upstream of the mapped location [118, 164, 178]. To ensure that differences in low quality bases would not bias our results, we treated any N in the genome annotation as an A. All raw read counts, mapped read counts, and filtered read counts can be found in Table 3.2.

We first identified an inclusive set of PAS in each species separately. We used the same in-house peak caller described in Mittleman *et al* [118], annotating each PAS as the most 3' base in each peak. The initial PAS set included 340,023 in human and 303,249 in chimps. We extended PAS 100 bp upstream and 100 bp downstream and used a reciprocal liftover pipeline to identify an inclusive set of orthologous PAS. We downloaded chain files from UCSC genome browser [78]. Details of the pipeline and number of PAS passing each step can be found in Supplementary Figure 3.29.

Due to gene annotation differences between species, we annotated all orthologous PAS to the human NCBI RefSeq annotation downloaded from UCSC genome browser [151], We used a hierarchical model to assign PAS to genic locations[101, 118]. We prioritized annotations in

the following order: 3' UTRs (UTR3), 5kb downstream of genes (end), exons (cds), 5' UTRs (UTR5), and introns (intron). We quantified reads mapping to each annotated PAS for each individual in both the total RNA libraries and nuclear RNA libraries using featureCounts with the -s strand specificity flag [98]. We calculated usage for each PAS in each library as a ratio of reads mapping to the PAS divided by the number of reads mapping to any PAS in the same gene (Supplementary Figure 3.8). We implemented two filtering steps to remove PAS with ratios likely biased by low site count or low gene count separately in each fraction.

Next, we filtered out sites with less than 5% usage in both species in the nuclear fraction. We then merged nuclear counts across all PAS in each gene. We removed PAS in genes not passing a cutoff of $log_2(CPM) > 2$ in at least 8 of the 12 individuals. After applying these filters, we were left with 44,432 PAS. As a quality control metric, we compared PAS usage calculated from the nuclear fraction to PAS usage calculated from whole cell fraction for each individual (we used the same methods to identify and quantify PAS usage in the whole cell 3' Seq data). We expected a high correlation between PAS usage in each fraction. Further, we expected a similar correlation in human and chimpanzee individuals [118]. Human individual NA18499 had significantly lower across-species correlation than the other individuals and was therefore removed from the analysis (Supplementary Figure 3.9.

To ensure gene expression level did not introduce ascertainment bias, we tested the relationship between PAS number and normalized gene expression. In both species, number of PAS is negatively correlated with normalized gene expression (Human: Pearson's correlation =-0.19, $p < 2.2x10^{-16}$, Chimp: Pearson's correlation=-0.17, $p < 2.2x10^{-16}$, Supplementary Figure 3.12). We expected species to contribute the most amount of variation to PAS usage. We ran PCA on the filtered nuclear PAS usage. PC1 accounts for 41.8% of the variation and is highly correlated with species ($R^2 = 0.68$). PC2 accounts for 13.1% of the variation and is moderately correlated with RNA extraction technician ($R^2 = 0.38$) and extraction day ($R^2 = 0.28$). As both of these variables are balanced with respect to species, we do

not believe they bias the results (Supplementary Figure 3.11). We identified 302 sites used at a rate of 5% in humans and 0% in chimpanzees, which we designated as human-specific. We identified 357 sites used at a rate of 5% in chimpanzee and 0% in humans, which we designated as chimp-specific.

We acknowledge the possibility that unlifted PAS may affect downstream analyses; therefore, we removed genes for which PAS ratios may be affected. Specifically, we annotated and calculated usage for the human PAS, including the 10,077 PAS that do not reciprocally lift to the chimpanzee genome. After removing PAS in genes previously identified as lowly expressed and PAS with usage below 5%, 386 PAS in 353 genes remain (Supplementary Figure 3.30). We removed these 353 genes and recreated Figures 3.3, 3.4, 3.5, 3.6 (Supplementary Figures 3.31, 3.32, 3.33, 3.34).

3.5.3 Orthologous 3' UTRs

We identified a set of orthologous UTRs using the ortho exon file described the Differential Expression section of the methods. We merged all regions annotated as 3' UTR by gene. If a gene had multiple non-continuous annotations, we selected the most 3' region as the orthologous UTR. We used deepTools compute matrix and plotHeatmap functions to plot merged human and chimpanzee reads along the orthologous 3' UTR set (Supplementary Figure 3.7, [154]) For all genes with PAS only in 3' UTRs, we assigned PAS to single, first, middle, and last, as previously described [190].

3.5.4 Analysis of sequence conservation around PAS

We used phyloP scores to measure sequence level conservation. We downloaded the hg38 100-way vertebrate PhyloP bigwig file from the UCSC table browser [147]. We computed scores for PAS regions as well as 200 bp intervals by taking the mean of the base pair scores. We removed any region with missing data from the analysis. We tested for differences in

mean phloP scores using Wilcoxon rank sum tests.

We tested for presence of the polyadenylation signal site motif in the 200 bp PAS regions. We used the bedtools nuc tool with the strand-specific flag to test for presence of each of the 12 previously annotated motifs for each PAS in both species [12, 152]. If a PAS had multiple motifs, we used a hierarchical model to choose the site based on the number of PAS with each identified motif (order: AATAAA, ATTAAA, AAAAAG, AAAAAA, TATAAA, AATATA, AGTAAA, AATACA, GATAAA, AATAGA, CATAAA, ACTAAA). The proportion of PAS with each signal site motif matched across species (Figure 3.1). To ask if presence or absence of a signal site explained species specificity or site-level differences, we restricted our analysis to the top two signal sites. These two motifs are the only sites where presence of a signal is associated with increased usage of the site in both species (Supplementary Figure 3.15). For the 359 PAS with one of these two signal sites present only in chimpanzees, average usage was higher in chimpanzees than in humans ($p = 0.025$). For the 361 PAS with one of these two signal sites present only in humans, average usage was higher in humans ($p = 2.0x10^{-4}$). We used hypergeometric tests to evaluate enrichment of differentially used PAS and species-specific PAS in the set of PAS with signal sites in only one species.

We also examined the proportion of U nucleotides in each PAS region. We used the bedtools nuc with the -s flag for strand specificity [152]. We tested if PAS with differences in U content are enriched for differentially used PAS using a hypergeometric test.

3.5.5 Differential APA

PAS level differences: We quantified reads mapping to each PAS using the featureCounts tool with the -s strand specificity tool [98]. We tested for site-level differences between human and chimpanzee using the leafcutter leafcutter_ds.R tool with standard settings [94]. We tested for differences in both the total and nuclear fractions. We tested 43,038 PAS in 8,422 genes in the nuclear fraction and 41,914 PAS in 8,333 genes in the total fraction. We

classified PAS as differentially used if the gene reached significance at 5% FDR and the PAS had a ΔPAU greater than 20% (absolute value ($\Delta PAU > 0.2$). A negative ΔPAU indicates increased usage in chimpanzees and ΔPAU indicates increased usage in humans. Top PAS per gene is the PAS with the most significant difference between species; ties were broken using mean usage for all individuals in both species.

Isoform diversity differences: We calculated Shannon Information content ($-\sum_{i=1}^{S} p_i log_2 p_i$) and Simpson's Index ($1 - \sum_{i=1}^{S} p_i^2$) using mean usage of each PAS in humans and chimpanzees, where p_i^2 is the usage of the i^{th} of s sites in the gene. We used Simpson's Index to assess isoform diversity because the correlation between Simpson's Index and number of PAS is lower than the correlation between Shannon Information content and the number of PAS per gene (Supplementary Figure 3.22, 3.23). To identify genes with differences in isoform diversity, we recalculated Simpson's index per gene per individual and tested for differences between species with Wilcoxon tests. We reported genes with differences at 5% FDR.

Conservation of dominant PAS: We consider a gene to have a dominant PAS if the within species average usage of the top used PAS is greater than the second most used site by an arbitrary cutoff. We reported results for cutoffs between 0.1 and 0.9. If a gene had a dominant PAS in either species, we included the top used site for both species when testing if genes use the same or different dominant PAS between species. We tested for enrichment of genes using the same or different dominant PAS with differentially expressed genes using hypergeometric tests.

3.5.6 Differential expression analysis

We generated unstranded RNA-seq libraries using the Illumina TruSeq Total RNA kit according to the manufacturer's instructions using the total mRNA collected from all 12 individuals (Illumina, San Diego, CA, USA). We sequenced RNA-seq libraries at the University of Chicago Genomics Core facility using the single-end 50 bp protocol on one lane of the

Illumina HiSeq 4000 machine. RNA quality and concentration at the time of library prep and number of sequenced reads per library are available in Table 3.3. We mapped the human libraries to GRhg38 [163] and chimpanzee libraries to panTro6 [33] and quantified reads mapping to orthologous exons.

To generate an updated orthologous exon file for the most recent chimpanzee genome assembly *(panTro6)* we followed the procedure reported in Pavlovic *et al.* with slight modifications [143]. We started with human *(GRCh38)* exon definitions from Ensembl version 98. We filtered this set of definitions for biotypes 'protein_coding,' using the command mkgtf from cellranger (10XGenomics). We then removed exon segments that were in exon definitions for multiple genes. This broke some exons into smaller unique exons. We then removed exons smaller than 10 bp. We took the final set of exons (1,371,917 exons from 20,338 genes) and extracted their sequences from the genome Ensembl GRCh38.p12. We used BLAT V. 35 to identify orthologous sequences within the chimpanzee genome *(panTro6)* [78] We removed hits with indels larger than 25 bp (using a function blatOutIndelIdent from `https://bitbucket.org/ee_reh_neh/orthoexon`). We then extracted the panTro6 sequences that had the highest sequence identity. We ran BLAT on this orthologous exon set to find matches in both the human and chimpanzee genomes. We removed exons that did not return the original location in humans or chimpanzees, as well as exons that mapped to multiple places with higher than 90% sequence identity. We removed exons that different human genes that mapped to overlapping regions in the chimpanzee genome. Finally, we removed exons that mapped to a different contig than the majority of exons from each gene. This resulted in a set of 1,250,820 orthologous exons from 19,515 genes.

We mapped on average 18.6 million reads to orthologous exons. We collapsed orthologous exons to quantify raw gene expression for each gene in each individual. We standardized counts and filtered out genes without $log_2(CPM)$ values greater than 1 in 8 of the 12 individuals. To prepare counts for differential expression modeling we used the Voom function

with the quantile normalization method in the limma R package [157]. We used PCA to test for batch effects. PC1 explains 35.1% of the variation and is highly correlated with species ($R^2 = 0.98$) (Supplementary Figure 3.35). Collected metadata such as the percent of live cells at collection, cell concentration at collection, RIN score, and RNA concentration do not segregate by species (Supplementary Figure 3.35). We modeled species as a fixed effect and called genes as differentially expressed at a 5% FDR. The results from our differential expression analysis, including effect sizes and significance values are available in Table 3.4.

3.5.7 Integration of translation and protein data

We downloaded differential translation genes and effect sizes from Additional file 5 of Wang *et al*[191]. Wang *et al.* modeled differential translation using ribosome profiling of 4 human, 4 chimpanzee, and 4 rhesus macaque LCLs. For all integrations, we conditioned on the 6,407 genes tested in the Wang *et al.* study and in our APA analysis. We tested for enrichments using a one-sided hypergeometric test implemented in R. We tested for correlations in effect sizes between site level ΔPAU and translation HvC effect sizes by first filtering for the top PAS (see top PAS method above, Supplementary Figure 3.26). We report Pearson's correlations calculated in R.

We downloaded differential protein level genes, effect sizes, and directional selection classifications from table S4 of Khan *et al.*[80]. Khan *et al.* modeled differential protein expression of 3,390 genes using high resolution mass spectrometry of stable isotope labeling by amino acids in cell culture *(SILAC)* collected from 5 human, 5 chimpanzee, and 5 rhesus macaque LCLs.

3.5.8 Supplemental functional data

We downloaded human protein length (in number of amino acids) for proteins annotated as reviewed for high confidence from UniProtKB [177]. We downloaded ubiquitination protein

modification data from PhosphoSitePlus version 050320 [65]. For all analyses in which we used interaction or ubiquitination data, we normalized the values by number of amino acids. To identify 3' UTR AU-rich elements in human RefSeq annotated 3' UTRs, we used the transcriptome_properties.py script published in Floor and Doudna 2016, available at `https://github.com/stephenfloor/tripseq-analysis`, with the -au-elements flag [51]. According to Floor and Doudna 2016, the fraction of AU-elements is the percentage of the 3' UTR with repeating AU elements of 5nt or more[51].

3.5.9 Data and code availability

All scripts and analysis pipelines can be found at `https://brimittleman.github.io/Comparative_APA/index.html`. FastQ files and PAS annotations are available on GEO under GSE155245.

3.6 Acknowledgments

We thank N. Gonzales for comments on the manuscript. We thank Y. Li, M. Ward, and G. Housman for useful discussion. Funding: This work was supported by the US National Institutes of Health (RO1 to Y.G). B.E.M. supported by T32 GM09197 to the University of Chicago and F31HL149259 to B.E.M. from National Heart, Lung, And Blood Institute of the National Institutes of Health. SP was in part supported by the National Center for Advancing Translational Sciences of the NIH (K12 HL119995). This work was completed in part with resources provided by the University of Chicago Research Computing Center.

3.7 Author Contributions

B.E.M. conceived the project with help from Y.G and S.P. B.E.M., S.W., C.C. and S.P. performed the experiments. B.E.M performed the analysis. K.B. curated orthologous exon

file. B.E.M. drafted the manuscript with input from Y.G. and S.P. S.P. and Y.G supervised the project.

3.8 Supplementary Information

3.8.1 Supplementary Figures

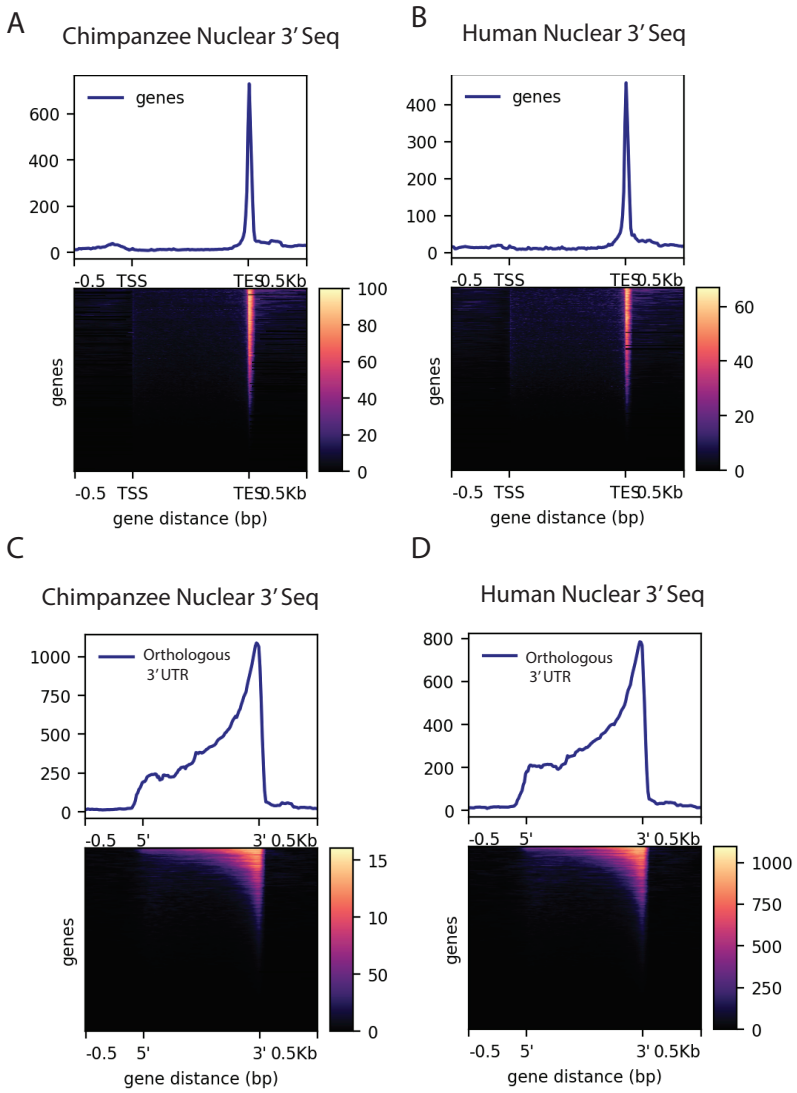


Figure 3.7: **Density of merged human and chimpanzee 3' Seq (A)** Coverage of 6 chimpanzee, nuclear 3' seq reads along Refseq transcripts **(B)** Coverage of 5 human, nuclear 3'seq reads along Refseq transcripts **(C)** Coverage of 6 chimpanzee, nuclear 3' seq reads orthologous 3' UTRs **(D)** Coverage of 5 human, nuclear 3' seq reads orthologous 3' UTRs

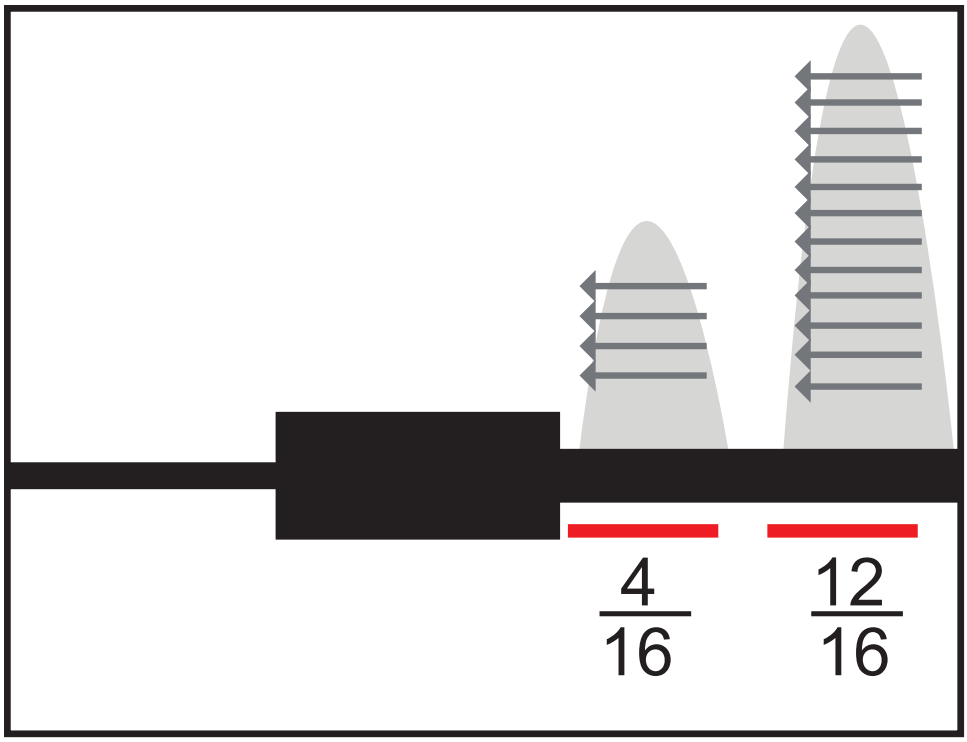


Figure 3.8: **Model representation of usage calculation** Representation of PAS usage calculation. Usage is a ratio of reads at each PAS to the number of reads mapping to any PAS in the same gene. Adapted from Chapter 2

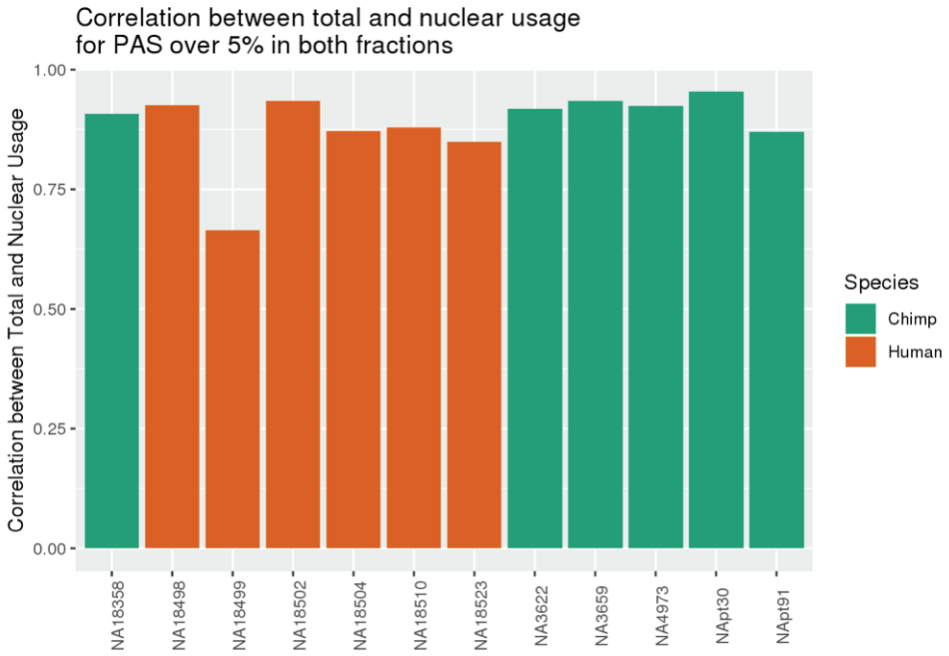


Figure 3.9: **NA18499 removed from analysis due to low correlation between fractions** Pearson's correlation between PAS usage calculated using nuclear 3' seq libraries and total mRNA 3' seq libraries, calculated using sites reaching 5% in one species in both fractions. (Plot from previous commit 30ff122 on Aril 9, 2020).

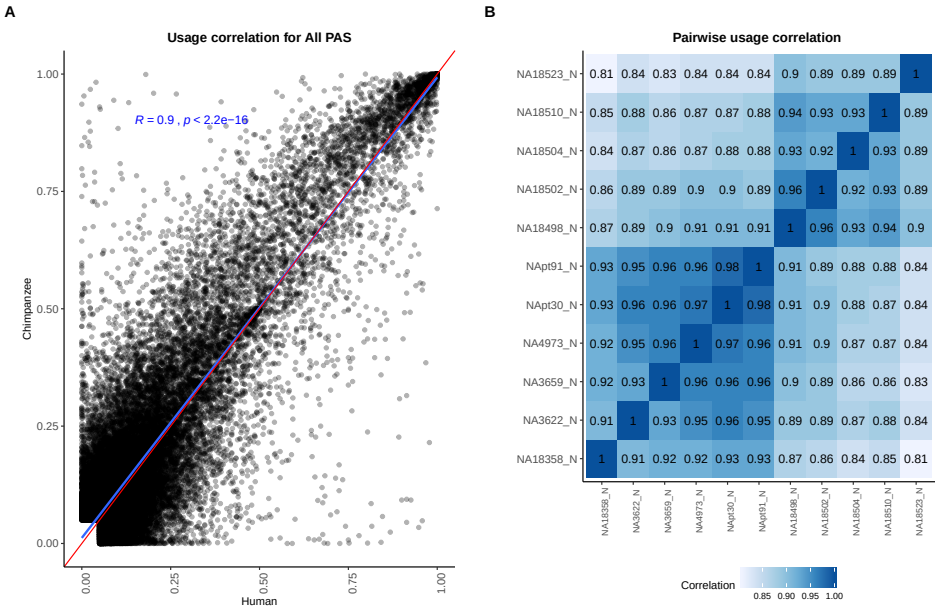


Figure 3.10: **PAS usage is highly correlated across species (A)** Correlation between human and chimpanzee PAS usage for 44,432 PAS. Red line is a 1:1 line. Linear regression line and Pearson's correlation plotted in blue. **(B)** Pairwise correlation for human and chimpanzee PAS usage.

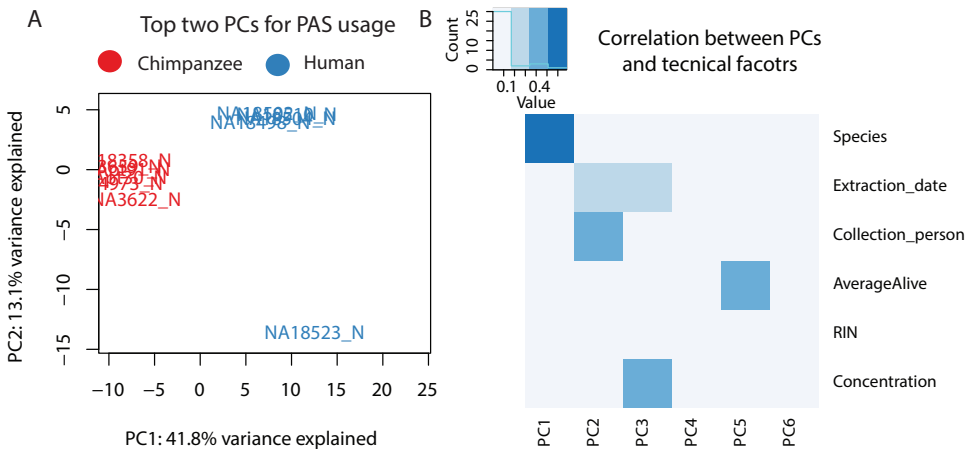


Figure 3.11: **Variation in PAS usage (A)** Plot of first two principal components *(PCs)* calculated by a principal component analysis on PAS usage (44,432 PAS). Chimpanzee samples are shown in red and human samples are shown in blue **(B)** Heatmap representing correlation between technical factors and PCs. Y axis factors include: Species, Extraction date, Collection person, AverageAlive (average of two live dead calculations at time of collection), RIN score, RNA concentration. Explanation of factors and values in Table 3.1

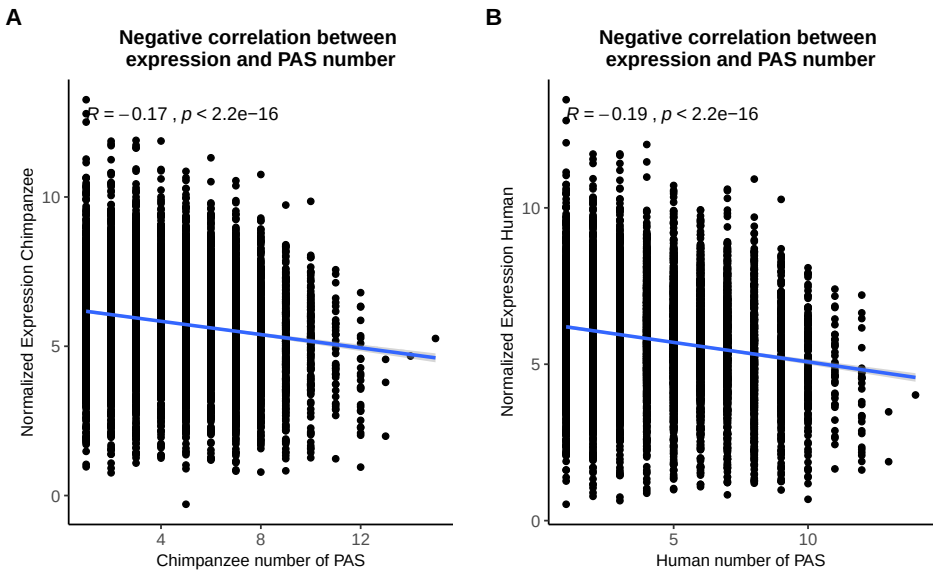


Figure 3.12: **PAS detection likely not biased by expression level (A)** Normalized gene expression plotted against the number of PAS detected at 5% usage in chimpanzee. **(B)** Normalized gene expression plotted against the number of PAS detected at 5% usage in human. The R package ggpubr was used to plot linear regression lines and calculate Pearson's correlations.

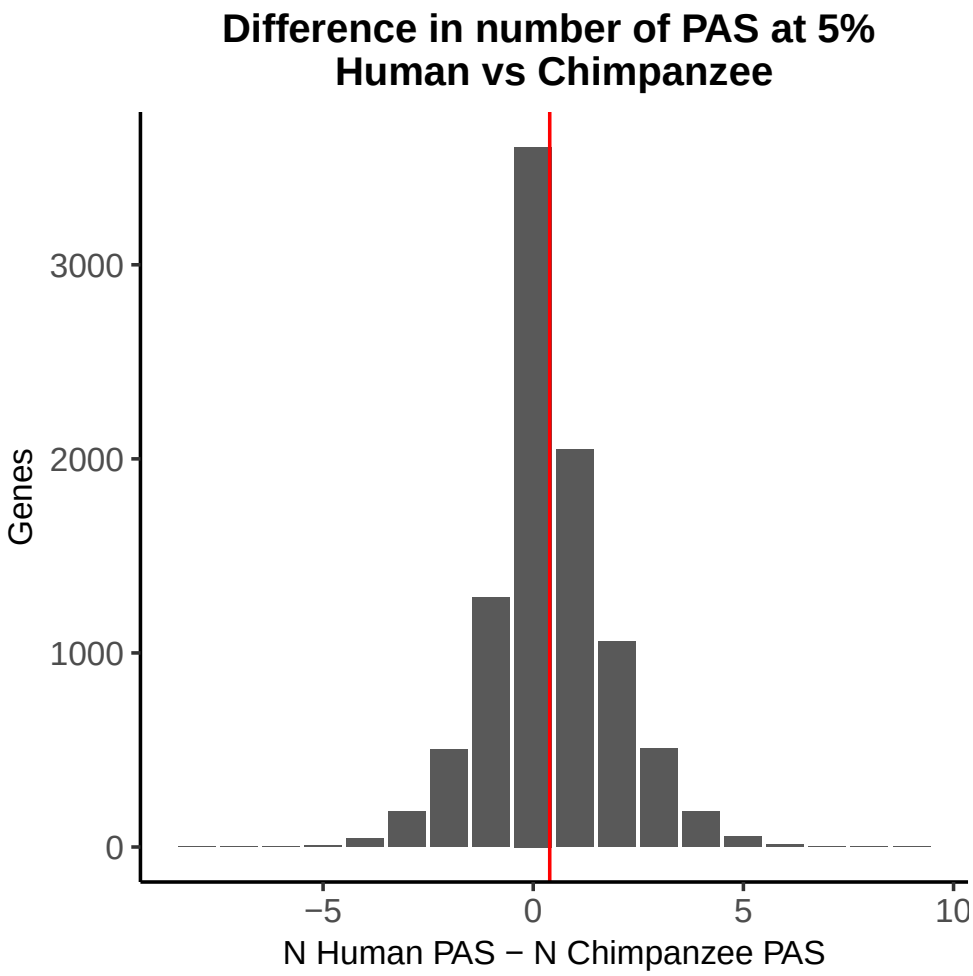


Figure 3.13: **PAS detection likely not biased by species** Histogram of the number of PAS detected at 5% usage in human minus the number of 5% usage in chimpanzees. Red vertical line represents mean difference (0.39).

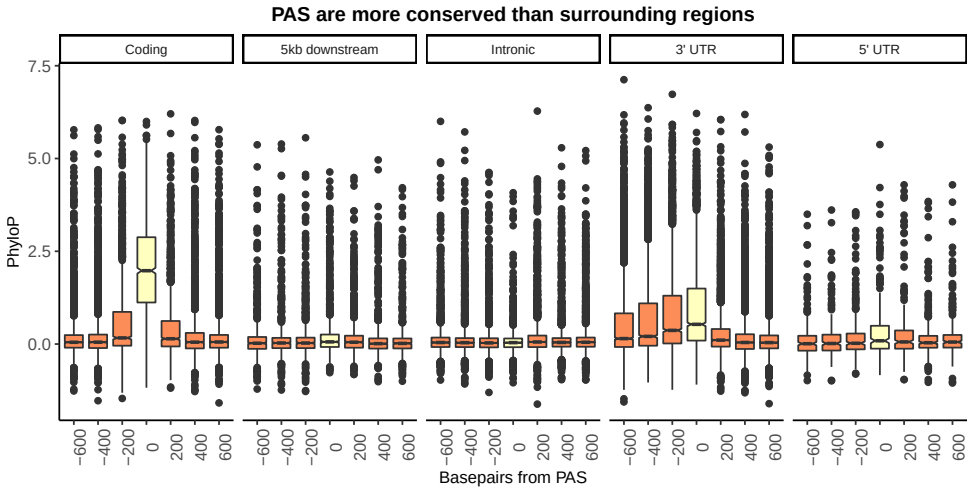


Figure 3.14: **Figure 1B separated by genic location** Mean PhyloP scores for PAS regions (yellow) and 200 base pair bins upstream and downstream of PAS (orange). A one-sided Wilcoxon test was used to test for increased PhyloP in PAS regions (Coding region: $p < 2.2x10^{-16}$, 5 kb downstream of genes: $p = 7.02x10^{-6}$, intron: $p = 0.99$, 3' UTR: $p < 2.2x10^{-16}$, 5' UTR: $p = 0.011$).

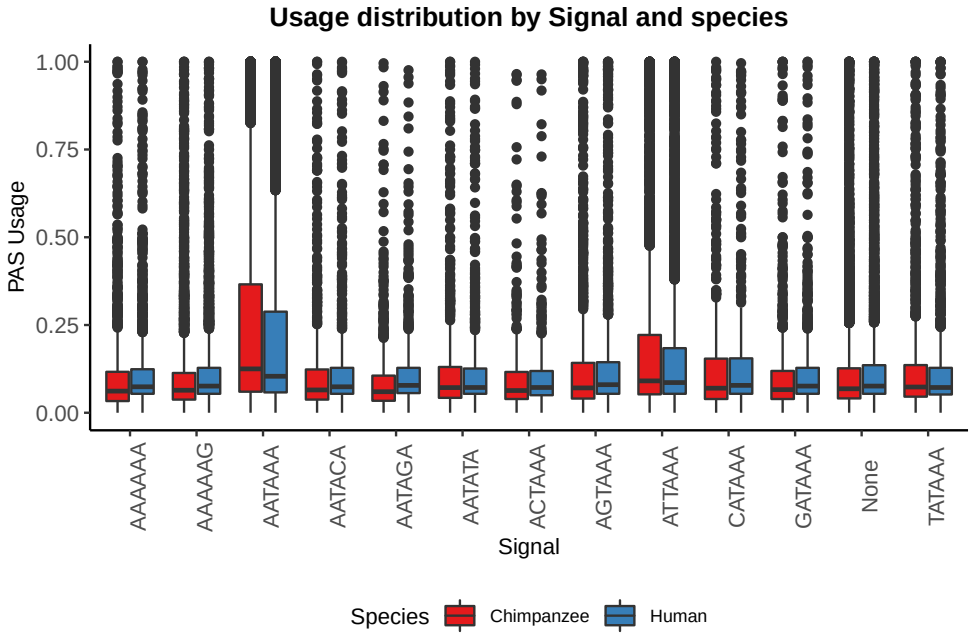


Figure 3.15: **PAS with AATAAA and ATTAAA are used more often** Mean PAS usage of the top two signal site motifs in human and chimpanzee plotted by annotated signal site.

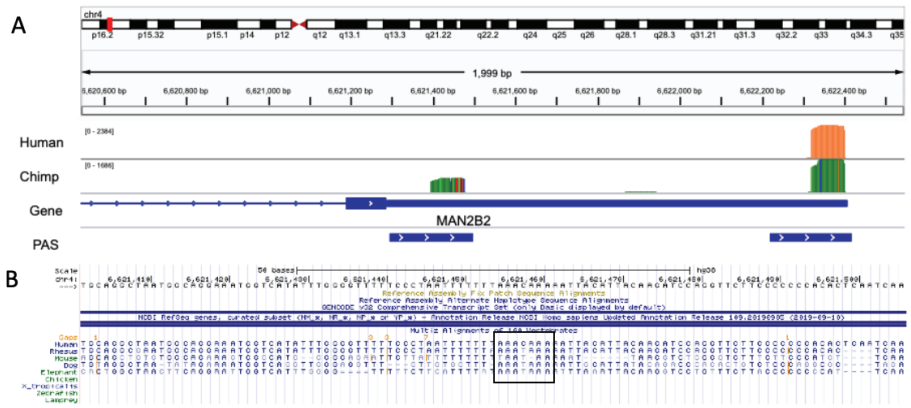


Figure 3.16: **Chimp specific PAS likely due to loss of signal site in human lineage (A)** IGV track for example of a chimpanzee specific PAS in MAN2B2 gene. Top track is merged coverage from 5 human nuclear 3' seq libraries. Chimp track is merged coverage from 6 chimpanzee nuclear 3' seq libraries lifted to human genome with CrossMap[206] **(B)** Sequence alignment for region upstream of proximal PAS from UCSC genome browser[78]. Black box indicates the signal site location. Canonical signal site is the ancestral state and was lost in the human lineage.

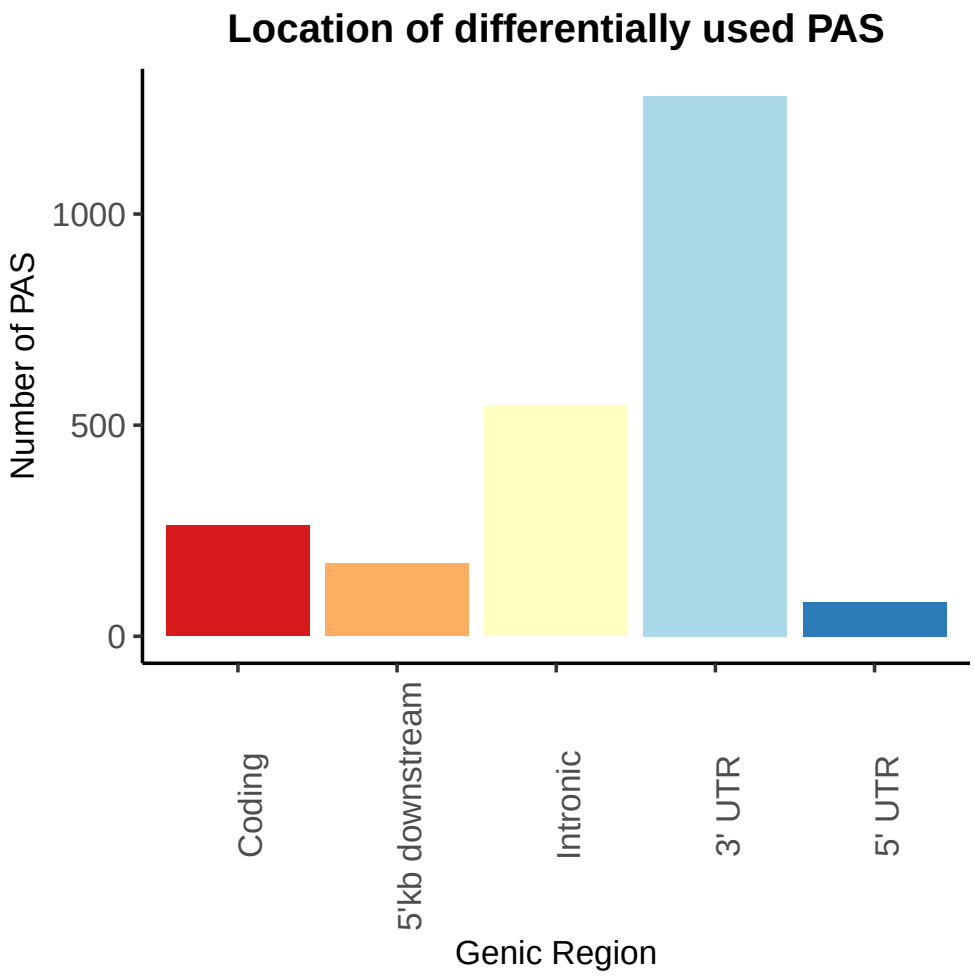


Figure 3.17: **Genic Location of PAS differentially used between human and chimpanzee** Differentially used PAS (5% FDR) between human chimpanzee by genic annotation.

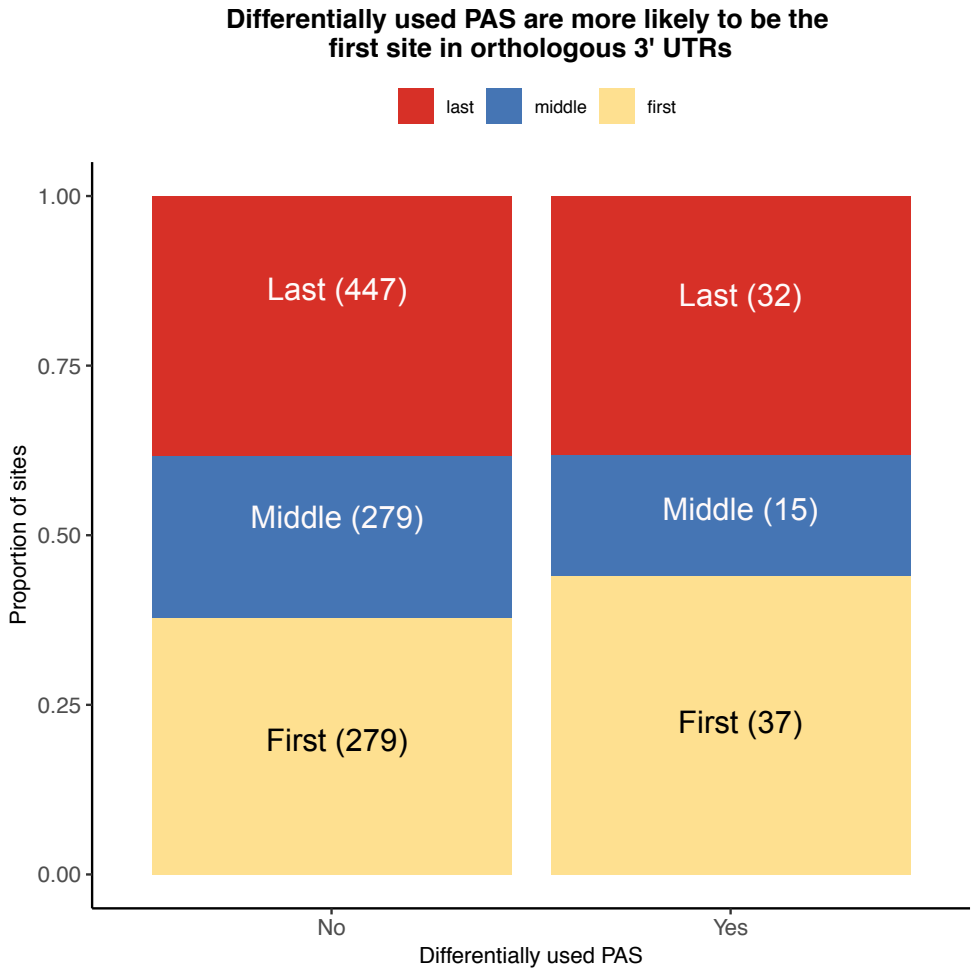


Figure 3.18: **Location of PAS within Orthologous 3' UTRs** Proportion of sites differentially used or conserved by whether they are the first (yellow), middle (blue), or last PAS (red) in orthologous exons.

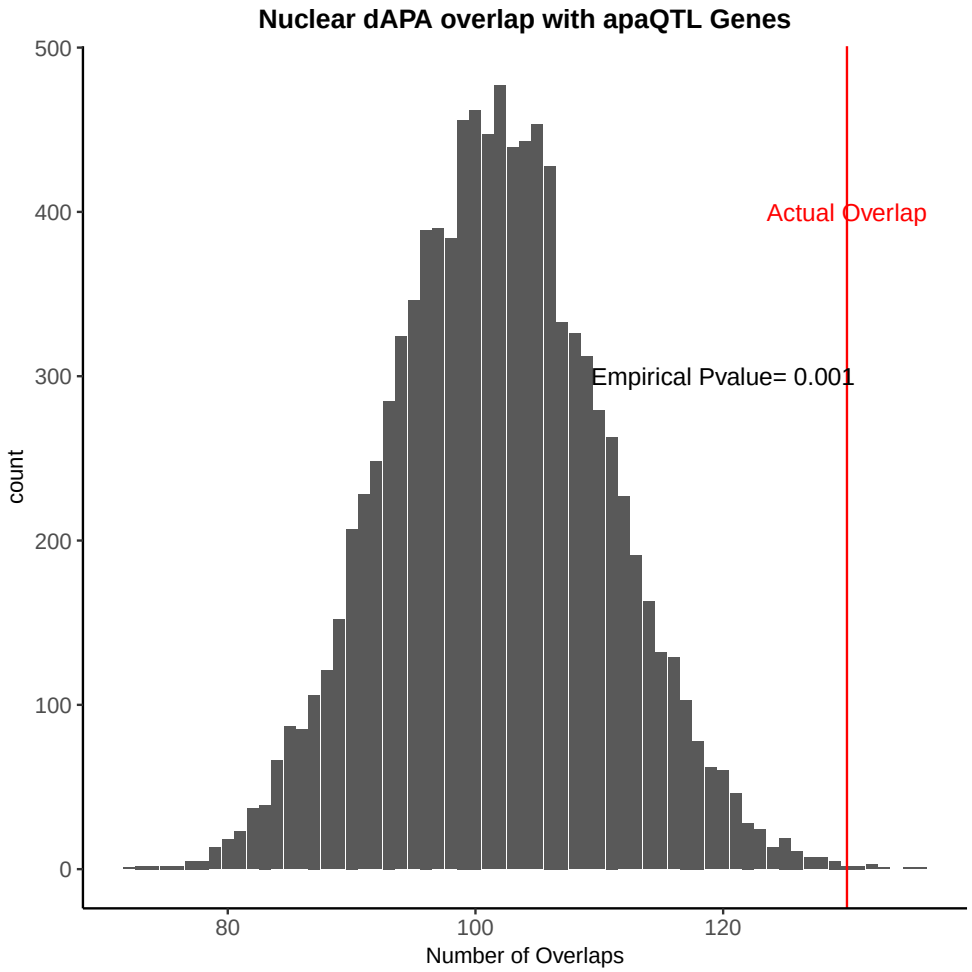


Figure 3.19: **Genes with differentially used PAS are enriched for genes with apaQTL** 10,000 random subsamples of genes tested for differential APA and overlap with genes with apaQTLs from Chapter 2. Red line represents the actual overlap between genes with differential usage of at least one PAS and apaQTL genes.

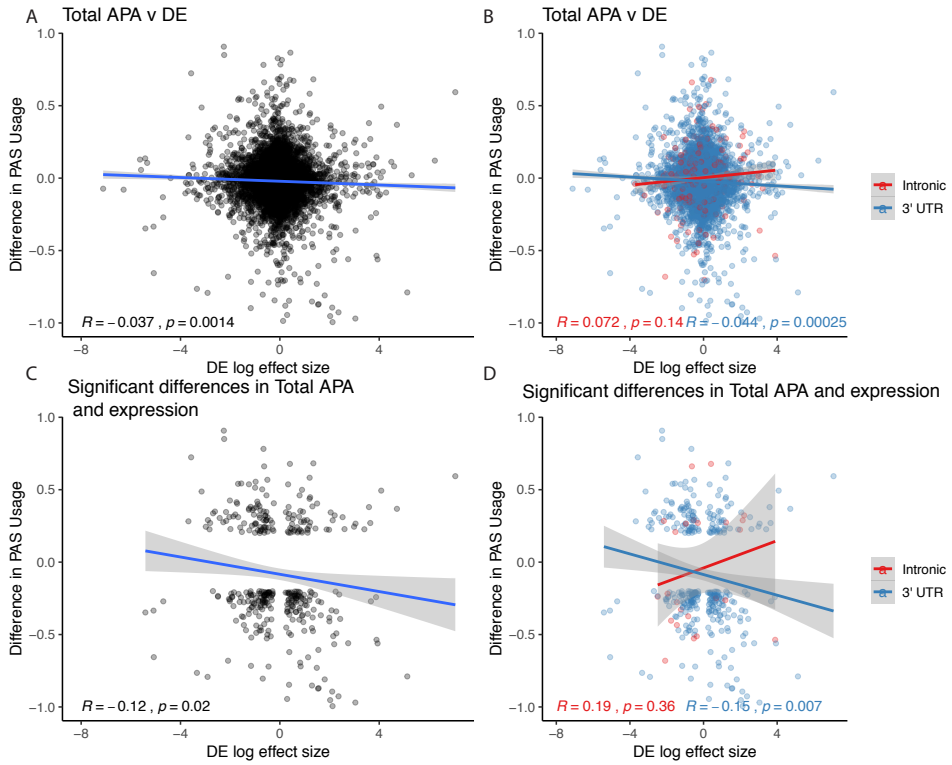


Figure 3.20: **Figure 3.3 relationships expanded to total usage (A)** Total mRNA ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis. **(B)** Total mRNA ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. **(C)** Total mRNA ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis. **(D)** Total mRNA ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. In all panels, I calculated the linear regression and Pearson's correlation with the r package ggpubr. In B and D, I colored the points and regression line by genic location. In all panels, negative ΔPAU and DE effect sizes represent upregulation in chimpanzees.

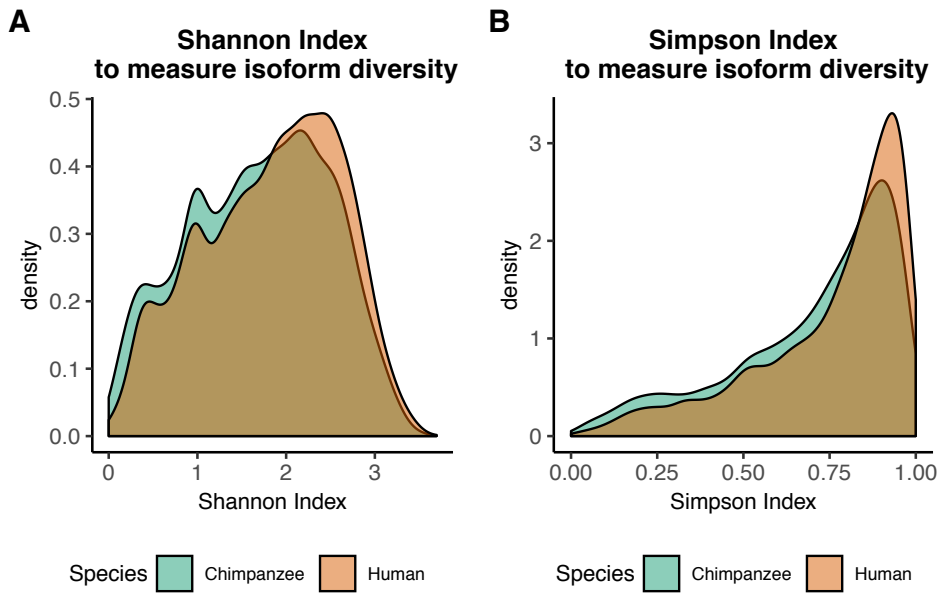


Figure 3.21: **Information content measurement densities (A)** Density of Shannon indices for all tested genes in human and chimpanzee ($-\sum_{i=1}^{S} p_i log_2 p_i$) **(B)** Density of Simpson indices for all tested genes in human and chimpanzee ($1 - \sum_{i=1}^{S} p_i^2$)

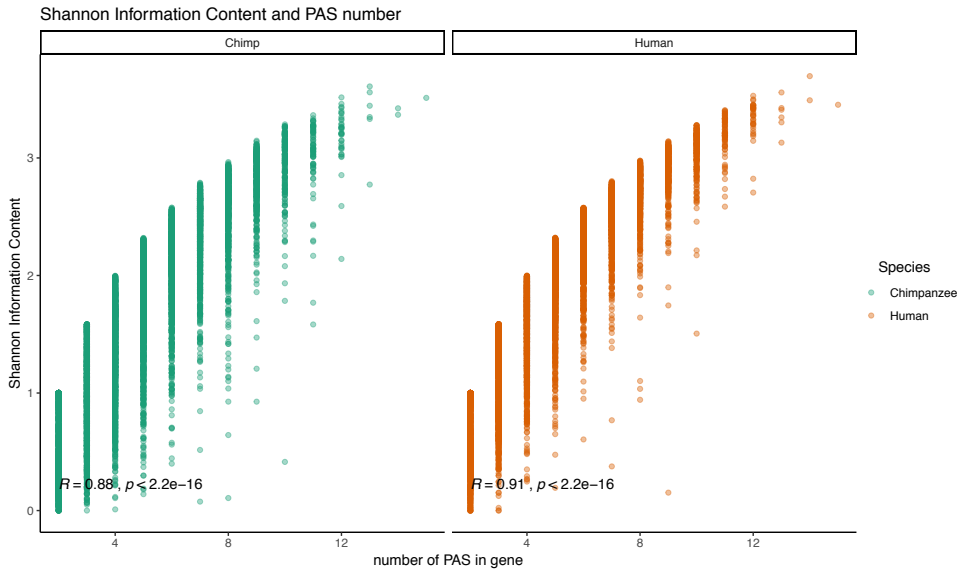


Figure 3.22: **Relationship between Shannon index and PAS number** Shannon information index plotted against the number of PAS detect for each gene. Pearson's correlation and significance in black.

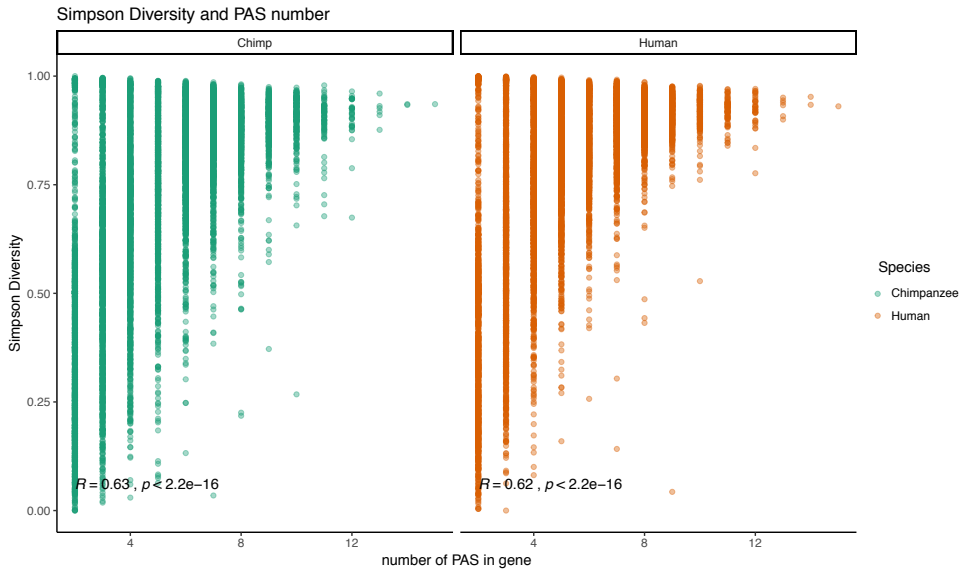


Figure 3.23: **Relationship between Simpson diversity index and PAS number** Simpson's diversity index plotted against the number of PAS detect for each gene. Pearson's correlation and significance in black.

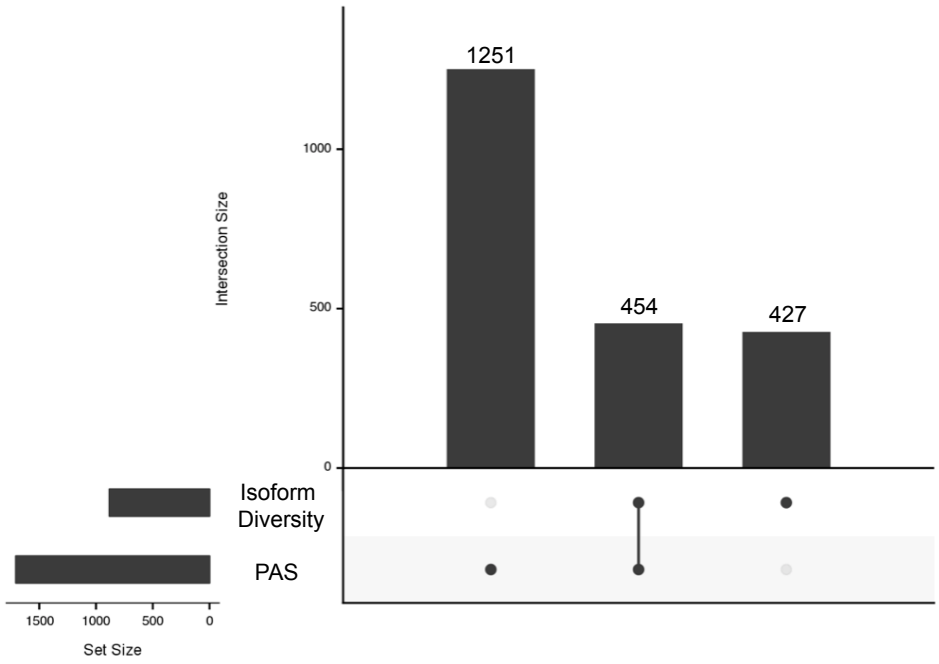


Figure 3.24: **Intersection between genes with PAS and isoform diversity differences** 1251 genes have significant differences in PAS usage at between human and chimpanzee (left). 454 genes have significant differences in APA between humans and chimpanzee in PAS usage and in isoform diversity (middle). 427 genes with differences in isoform diversity level only (right).

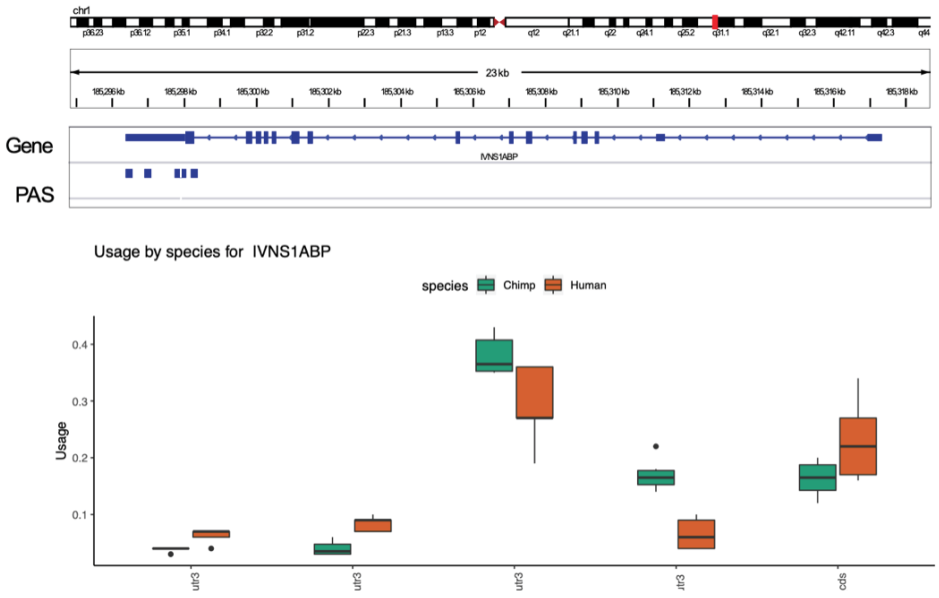


Figure 3.25: **Gene with significant differences in isoform diversity only** Human and chimpanzee usage for 5 PAS identified in the IVNS1ABP gene. None of the PAS measured have significant differences in usage at 5% FDR. IVNS1ABP is not differentially expressed.

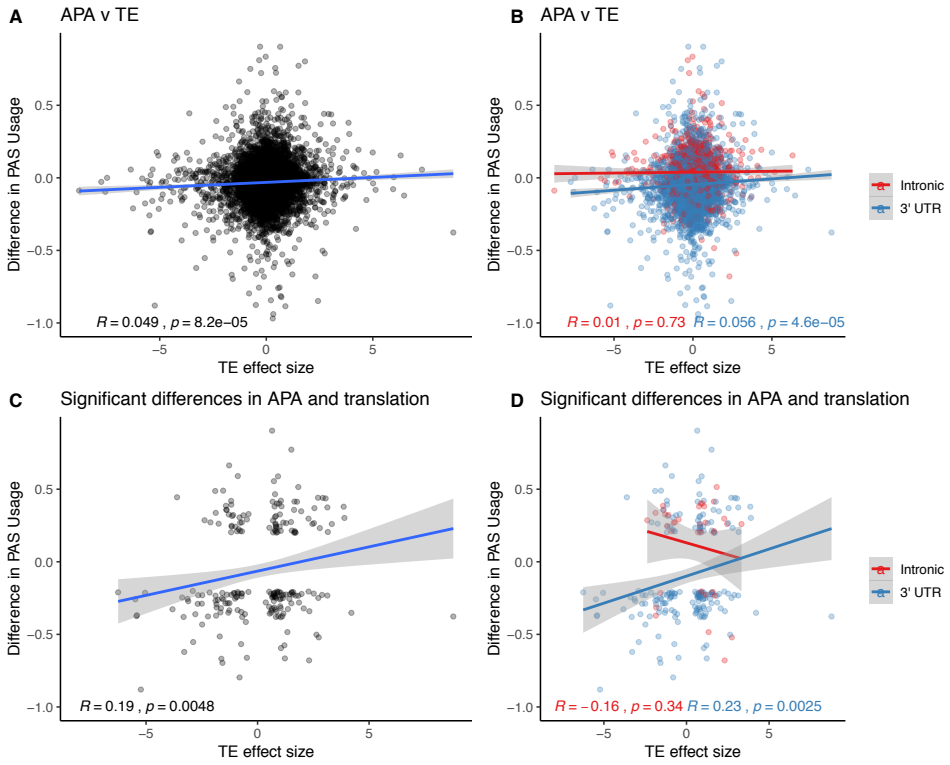


Figure 3.26: **Relationship between ΔPAU and differential translation effect sizes (A)** ΔPAU for top 3' UTR and intronic PAS plotted against differential translation *(TE)* effect size as reported by Wang *et al.*[191] **(B)** ΔPAU for top 3' UTR and intronic PAS plotted against TE effect size as reported by Wang *et al.* [191] separated by genic location. **(C)** ΔPAU for top 3' UTR and intronic PAS with significant differences in usage plotted against TE effect size for significant genes (5% FWER) as reported by Wang *et al.*[191]. **(D)** ΔPAU for top 3' UTR and intronic PAS with significant differences in usage plotted against TE effect size for significant genes (5% FWER) as reported by Wang *et al.*[191] separated by genic location. Linear regression line was plotted and Pearson's correlation was calculated for data in each panel.

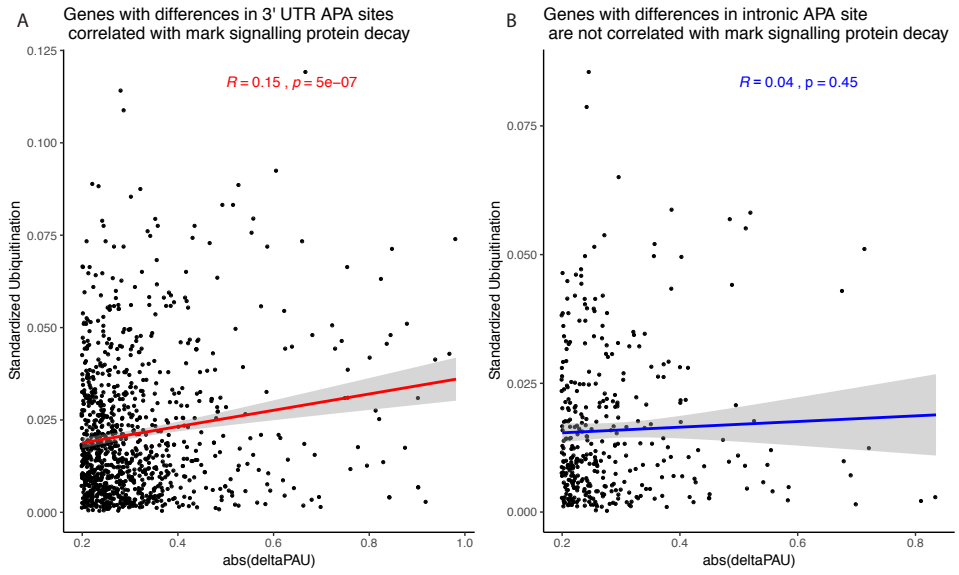


Figure 3.27: **Relationship between APA differences and protein decay mark (A)** Absolute value of ΔPAU for 3' UTR PAS with significant difference at site level plotted against the number of ubiquitination marks in the gene standardized by the number of amino acids. Regression line and Pearson's correlation are plotted in red.**(B)** Absolute value of ΔPAU for intronic PAS with significant difference at site level plotted against the number of ubiquitination marks in the gene standardized by the number of amino acids. Regression line and Pearson's correlation are plotted in blue.

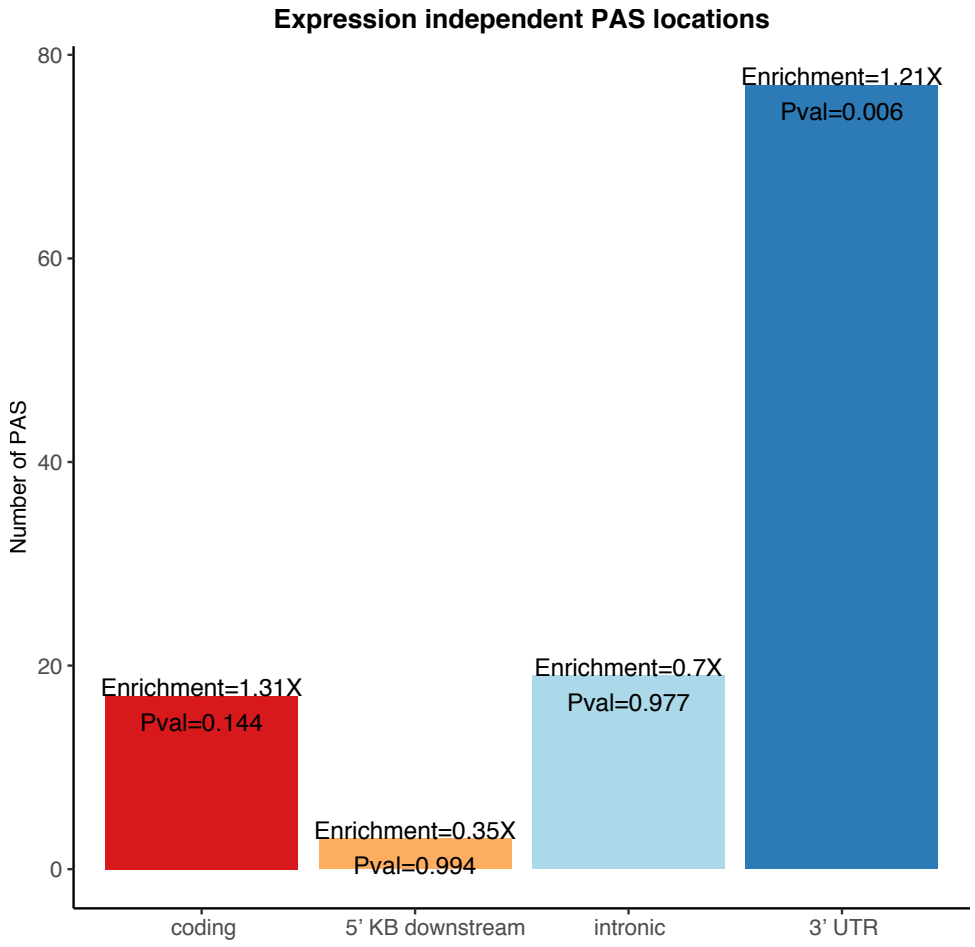


Figure 3.28: **Enrichment for 3' UTR PAS in genes differentially expressed in protein and not in mRNA** Genic location enrichments for the PAS in genes differentially expressed at protein level but not mRNA level among all differentially used PAS. Pvalues were calculated with a hypergeometric test.

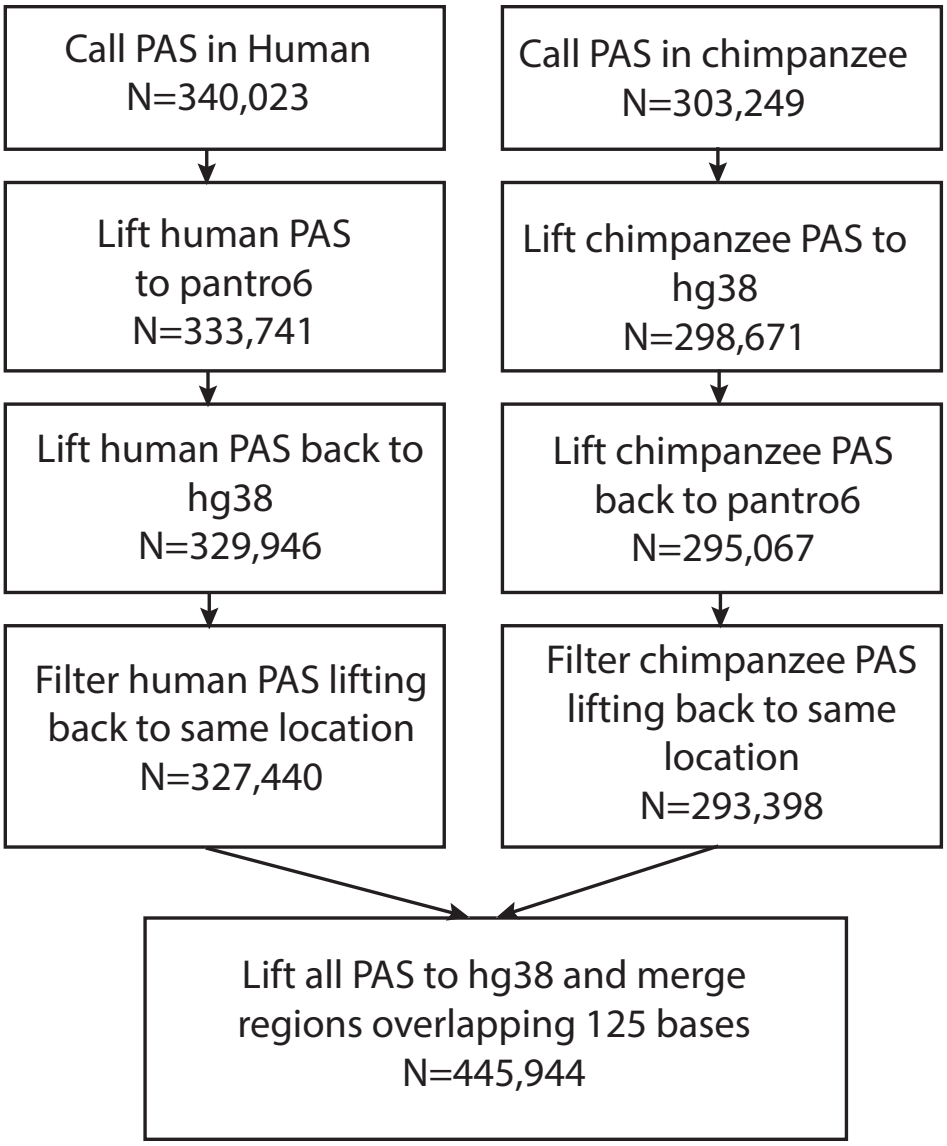


Figure 3.29: **Reciprocal liftover pipeline** Reciprocal liftover pipeline for unfiltered PAS including the number of sites remaining at each step. Liftover using UCSC liftover tool and chain files downloaded from UCSC genome browser [78].

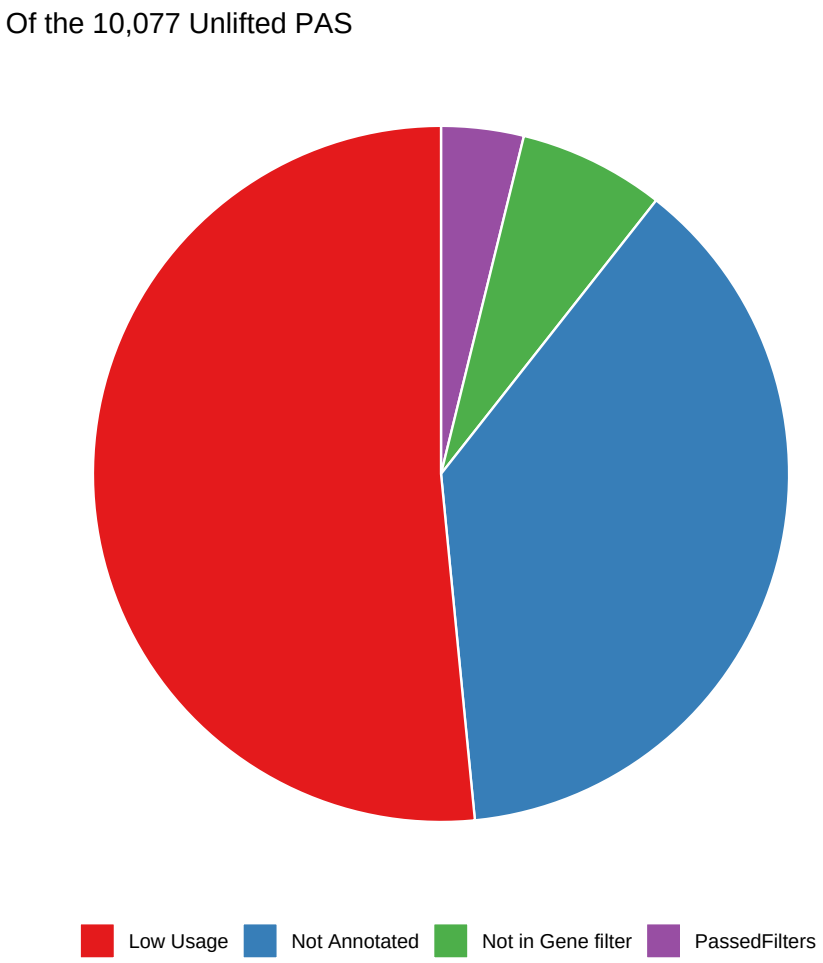


Figure 3.30: **PAS that do not lift from human to chimp** Of the 10,077 PAS that do not reciprocally lift from human to chimp, distribution of where sites are filtered out. Most are lost due to not mapping to genes or due to low usage (likely noise).

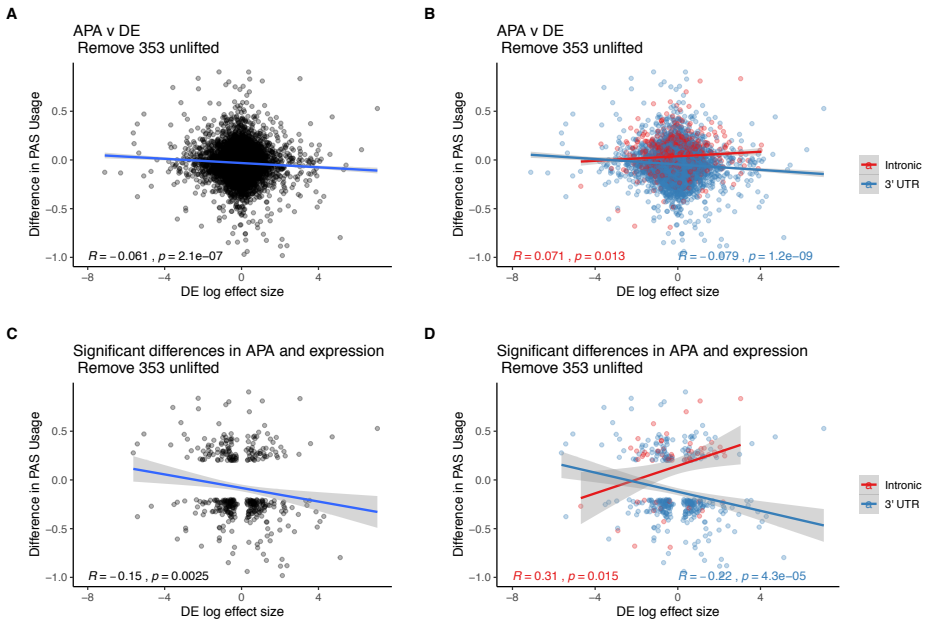


Figure 3.31: **Figure 3.3 without genes affected by liftover (A)** ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis. **(B)** ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. **(C)** ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis. **(D)** ΔPAU for top intronic or 3' UTR PAS per gene plotted against differential effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. In all panels, I calculated the linear regression and Pearson's correlation with the r package ggpubr. In B and D, I colored the points and regression line by genic location. In all panels, negative ΔPAU and DE effect sizes represent upregulation in chimpanzees.

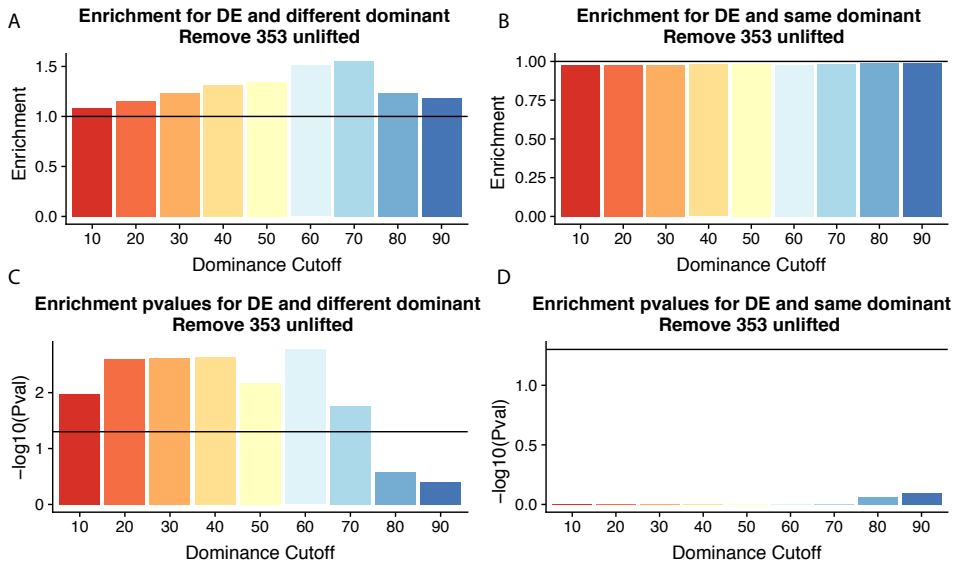


Figure 3.32: **Figure 3.4 without genes affected by liftover (A)** Enrichment of genes with the different (left) or same (right) dominant PAS by dominant cutoff in differentially expressed genes after removing genes likely affected by liftover. **(B)** $-log_{10}(p - values)$ for enrichments in A calculated with hypergeometric tests. Horizontal line represents $p = 0.05$.

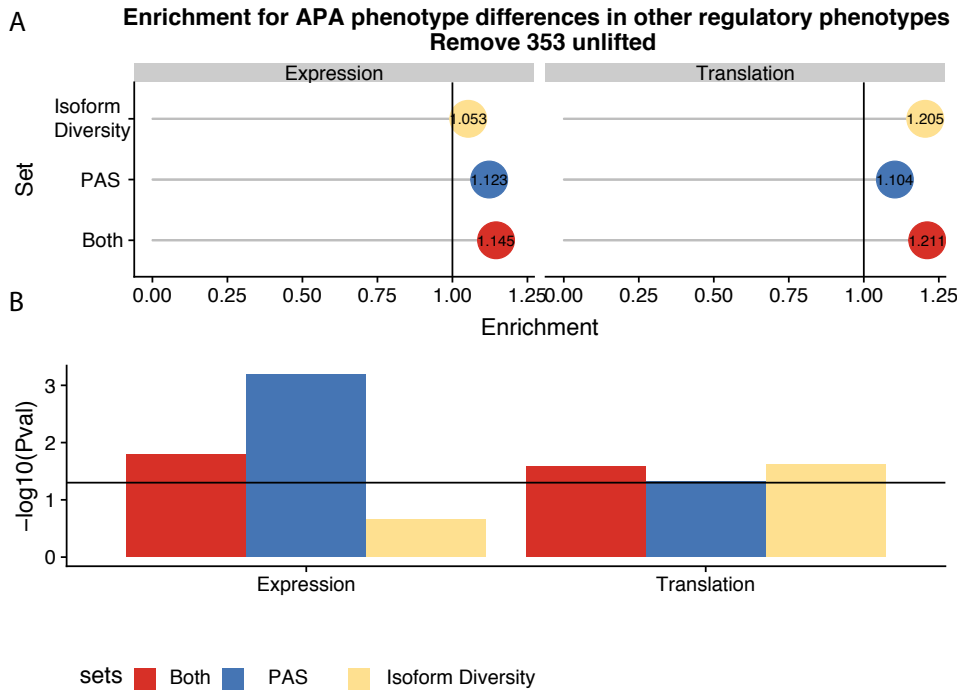


Figure 3.33: **Figure 3.5 without genes affected by liftover (A)** Enrichment of genes with differences in isoform diversity, PAS usage, or both within differential expressed genes and differentially translated genes after removing genes likely affected by liftover. Differentially translated genes reported by Wang *et al.*[191]. **(B)** $-log_{10}(p-values)$ for enrichments in A calculated with hypergeometric tests. Horizontal line represents $p = 0.05$.

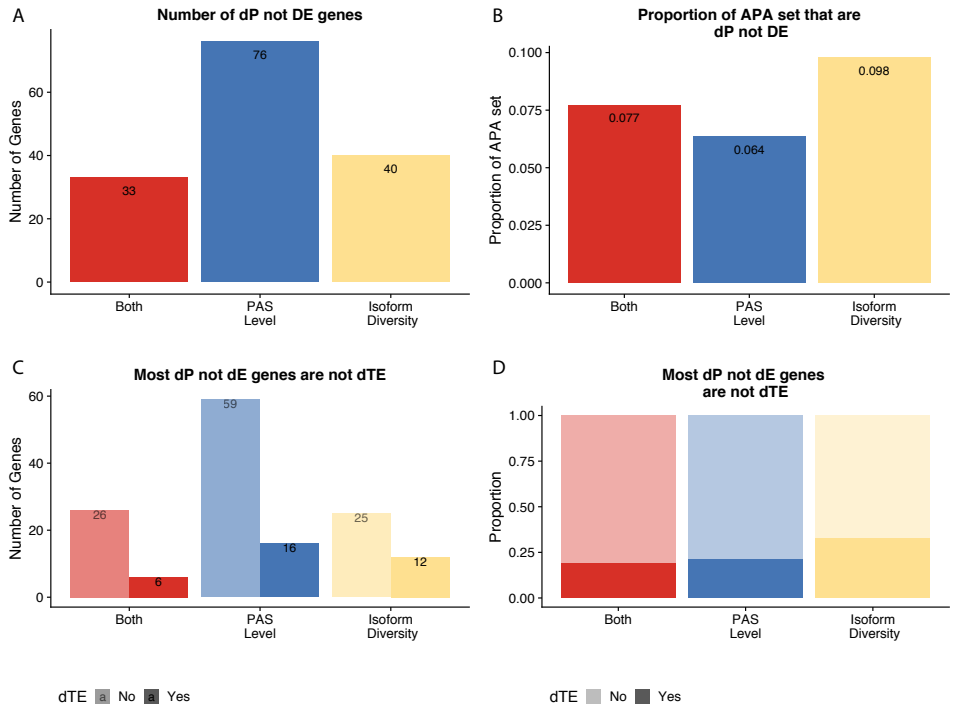


Figure 3.34: **Figure 3.6 without genes affected by liftover (A)** Number of genes with differences in isoform diversity, PAS usage or both differentially expressed in protein (5% FDR) but not in mRNA (5% FDR). Genes differentially expressed in protein from Khan *et al.*[80] **(B)** Proportion of genes with differential isoform diversity, PAS usage or both that are differentially expressed in protein (5% FDR), but not mRNA (5% FDR). **(C)** Genes reported in separated by genes differentially translated at 5% FDR. Differentially translated gene reported in Wang *et al.*[191]. **(D)** Genes differentially expressed in protein but not in mRNA, colored by differences in APA. Proportion of genes in the set differentially translated at 5% FDR.

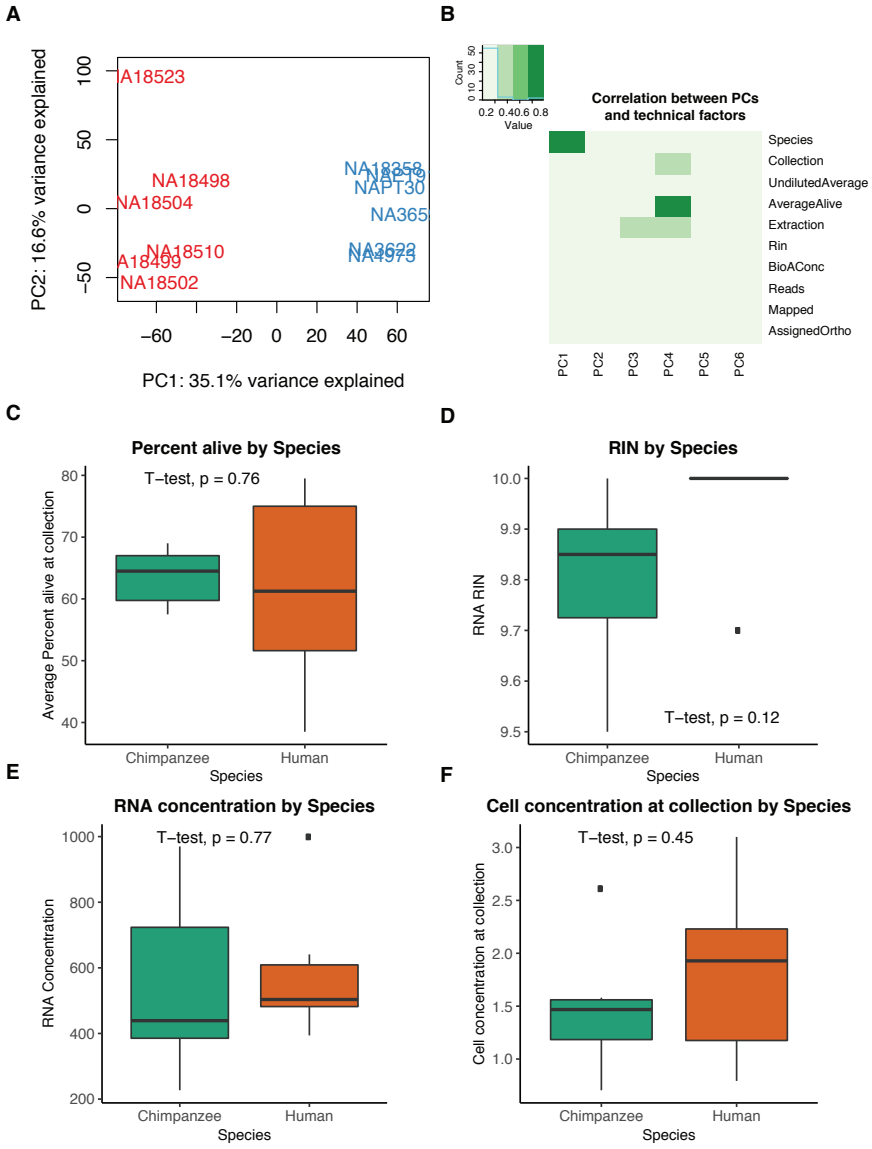


Figure 3.35: **Differential expression quality control plots (A)** First two principal components *(PCs)* in gene expression variation. **(B)** Heatmap representing correlation between technical factors and PCs. Explanation of Y axis factors and values available in Table 3.2 **(C)** Percent of live cells as calculated by trypan blue staining at collection is not confounded by species. **(D)** RIN scores reported by bioanalyzer at RNA-seq library generation are not confounded by species.**(E)** RNA concentrations reported by bioanalyzer at RNA-seq library generation are not confounded by species. **(F)** Cell concentrations at time of collection are not confounded by species.

3.9 Supplementary Tables

Table 3.1: **PAS differential usage** (see supplementary file associated with this dissertation) Differential PAS usage results. Column names as described- PAS: polyadenylation site, gene: gene (cluster in leafcutter), PAS_logeffectsize: log effect size for differential usage of the PAS between human and chimpanzee, PAS_deltaPAU: difference in polyadenylation site usage between human and chimpanzee (delta PSI in leafcutter), Gene_lohLR: log likelihood ratio for differential usage of any PAS in the gene. (cluster likelihood ratio in leafcutter), Gene_adjustedPvalue: adjusted pvalue for differential usage of the any PAS in the gene (cluster pvalue in leafcutter)

Table 3.2: **3' Seq metadata** (see supplementary file associated with this dissertation) Metadata for 3' Seq data. Column names as described- Species:Cell line species , Lines: Cell line ID, Fraction: Cellular fraction, CollectionDate: Date of cell harvest and nuclear isolation, Extraction_date: Date of RNA extraction, Collection_person: Author initial for who processed cell harvest and nuclear isolation, UndilutedAverage: Average of 2 cell count measurements $1x10^6$, AverageAlive: Average of 2 cell live dead counts - calculated with trypan blue stain, Concentration: Extracted RNA concentration (ng/ml), RIN: RIN score for extracted RNA, 260.280.Ratio: 260/280 ratio calculated on nanodrop, Library: 3' Seq library date, Reads: Number of sequenced reads, Mapped_wMP: Number of Mapped reads before removing reads likely due to misprimming, Mapped_Clean: Number of Mapped reads after removing reads likely due to misprimming

Table 3.3: **RNA sequencing metadata** (see supplementary file associated with this dissertation) Metadata for RNA sequencing data. Column names as described- Species:Cell line species , Lines: Cell line ID, Collection_person: Author initial for who processed cell harvest and nuclear isolation, UndilutedAverage: Average of 2 cell count measurements $1x10^6$, AverageAlive: Average of 2 cell live dead counts - calculated with trypan blue stain, CollectionDate: Date of cell harvest and nuclear isolation, Extraction: Date of RNA extraction, RIN: RIN score for extracted RNA, BioAConc: RNA concentration (ng/ul), Reads: Number of Sequenced reads, Mapped: Number of mapped reads, AssignedOrtho: Number of mapped reads assigned to orthologous exons.

Table 3.4: **Differential expression results** (see supplementary file associated with this dissertation) Differential expression results from limma. gene: tested gene, logFC: log 2 fold change in normalized gene expression, adj.P.Val: BH adjusted pvalue from t test, B: Beta value, t: t statistic

CHAPTER 4

NATIVE ELONGATING TRANSCRIPT SEQUENCING TO MEASURE POLYMERASE II ELONGATION RATE IN A HUMAN POPULATION

4.1 Abstract

In chapter 4, I describe a project in which we aimed to use Native Elongating transcript sequencing *(NET-seq)* to quantify polymerase II *(Pol II)* elongation speed variation genome wide in a panel of YRI LCLs. Our goal was to map genetic variation associated with Pol II elongation speed. We would then ask if these genetic variants were also correlated with previously identified regulatory phenotypes, such as gene expression and alternative splicing. Unfortunately, the NET-seq data was not of high enough quality or complexity to continue the analysis. While this work will not be published elsewhere, the work contributed to my development as a scientist and is thus included here. In this chapter, I will describe our motivation, efforts made, and suggest alternative approaches that may allow for the detection of genetic variation association Pol II pausing.

4.2 Introduction

Functional regulatory QTL studies have successfully uncovered a large number of interacting regulatory mechanisms likely responsible for variation in gene expression within human populations [95, 117, 42, 54, 10]. Such studies have primarily focused on identification of pre-transcriptional gene regulatory features, such as enhancers and promoters, through characterization of chromatin accessibility and histone modifications. However, many of the genetic variants correlated with variation in gene expression fall outside of promoter and enhancer regions.

Through a meta-analysis of previously characterized molecular phenotypes in the same panel of human lymphoblastoid cell lines *(LCL)* Li *et al.* quantified the proportion of expression quantitative trait loci *(eQTLs)* with a suggested molecular mechanism. Namely, Li *et al.*, estimated that 60% of the eQTLs previously identified in LCLs, likely contribute to differences in expression through variation at chromatin level features. This analysis left around 40% of eQTLs mechanistically unexplained [95]. The unexplained variants lie within gene bodies and were associated with regions of active transcription elongation, suggesting they act through co-transcriptional mechanisms.

It is likely that genetic variation associated with co-transcriptional mechanisms also contribute to isoform specific gene regulation. mRNA isoform variation arises through alternative splicing and alternative polyadenylation. Genetic variant associated with alternative splicing *(sQTLs)* and alternative polyadenylation *(apaQTLs)* are also likely driven by co-transcriptional gene regulation [95, 118]. In turn, a more thorough characterization of co-transcriptional gene regulatory mechanisms could improve our understanding of eQTL, sQTLs, and apaQTLs.

Using estimates of nascent transcription and polymerase II *(Pol II)* density, researchers have discovered that Pol II moves along gene bodies at a non-uniform rate [112, 127, 106, 39, 74]. Specifically, Pol II density increases proximal to the promoter, at intron exon

boundaries, and at the transcription end site *(TES)* suggesting Pol II pauses at each of these locations during transcription [1, 205, 153]. According to studies in human and other model systems, Pol II pausing is tightly regulated [127, 29, 149, 60]. While, various studies have mechanistically implicated Pol II dynamics in alternative splicing and APA, there is still debate surrounding causal relationships and the degree to which Pol II pauses or simply slows down [148, 156]. Moreover, despite the large body of molecular work, no study has quantified interindividual variation in Pol II elongation rate.

We suspect genetic variation contributing to differences in Pol II elongation rate are also associated with differences in gene expression, alternative splicing, and alternative polyadenylation. By identifying these genetic variants (pauseQTLs) we can expand our knowledge of gene regulatory mechanisms. We collected Native Elongation Transcript sequencing (NET-seq) data from a population of human lymphoblastoid cell lines *(LCLs)* in order to quantify variation in Pol II density as an estimate of Pol II elongation rates. We intended to map genetic variation associated with elongation differences to ask if Pol II elongation rate is a co-transcriptional gene regulatory mechanism contributing to variation in gene expression, alternative splicing, and alternative polyadenylation. Unfortunately, this project was not completed as intended because the NET-seq data was not complex enough to assess individual level variation genome wide.

4.3 Results

A number of protocols have been developed to measure Pol II density and nascent transcription genome wide [195]. For this project, we decided to use the Nascent Elongating Sequencing *(NET-seq)*protocol published by Andreas Mayer and L. Stirling Churchill in 2016 [111]. The protocol maps Pol II density genome wide at single-nucleotide precision without cell perturbation or nascent RNA labeling.

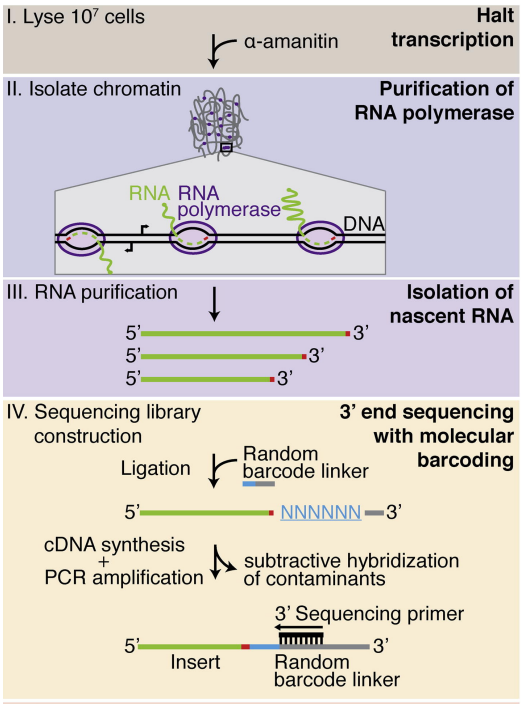
We optimized NET-seq for 16 human lymphoblastoid cell lines *(LCLs)*. First, we halted

transcription with α-Amanitin and purified nascent mRNA molecules from purified chromatin (Figure 4.1, Panel I-III). We added a DNA linker with a 6 base pair unique molecule identifier *(UMI)* to the 3' hydroxyl group of each nascent molecule. This step allows for base pair and strand specific detection of individual molecules during downstream bioinformatic steps. After using a gel extraction to select 30-100 nucleotide RNA fragments, we created circular cDNA from each fragment. Because some mature mRNAs, such as snoRNAs remain associated with chromatin and likely contribute to the cDNA pool, we used biotinylated DNA oligos to specifically deplete a number of previously annotated, snRNAs, snoRNAs, rRNAs, and mitochondrial tRNAs. The remaining set of cDNA's required a fairly high number of PCR amplification cycles (12-20) to achieve libraries with concentrations high enough to sequence. (Figure 4.1, Panel IV).

We sequenced the libraries to an average depth of 160 million reads. For many libraries, less than 50% of the reads mapped to the genome and once deduplicated based on UMIs, less than 5% of sequenced reads were usable. Over 50% of reads did not map because they were too short. Mapped reads from our libraries were shorter reads than those previously published ([112], Figure 4.2A-D)

We next evaluated the mapped data on a genome wide scale. We assessed our data using the following metrics introduced by Meyer *et al.* (mayer 2015). At the gene coverage level, NET-seq libraries were highly correlated (Figure 4.2E). Overall, we observed a bias toward read coverage at the 5' end of gene bodies (Methods, Figure 4.3A). Within genes, we observed enrichment at 5' and 3' exon boundaries for the top 5% of expressed exons (Figure 4.3B, Methods). After standardizing the number of reads mapping to each gene by gene length only, on average 42.7% of genes were detected at greater than 0.001 standardized reads (Figure 4.3C). Given the average gene length, 0.001 standardize reads represented about 65 reads.

Our goal was to identify genomic regions with evidence for high Pol II density and map



Adopted from Mayer et al. 2015, figure 1

Figure 4.1: **Graphical representation of NET-seq protocol published in Mayer *et al.* [112] Panel I:** Halt transcription in cells with α-Amantin. **Panel II:** Purify chromatin containing Pol II. **Panel III:** Purify nascent mRNA from chromatin fractionation. **Panel IV:** Library construction by adding DNA linker to 3' OH group, cDNA synthesis, and removal of mature mRNA contaminates

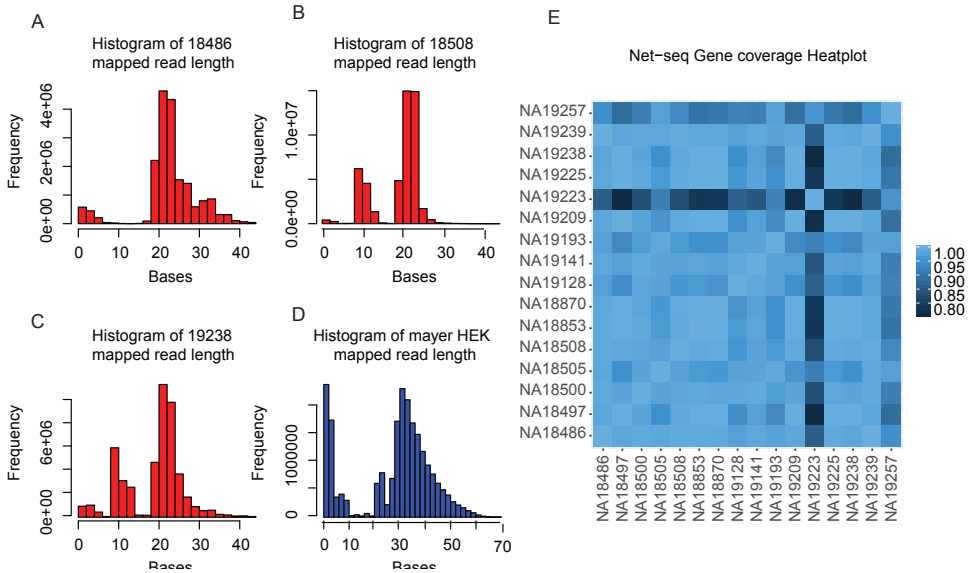


Figure 4.2: **Quality control metrics for NET-seq libraries. A** Histogram of mapped read lengths for NET-seq library NA18486. **B** Histogram of mapped read lengths for NET-seq library NA18508. **C** Histogram of mapped read lengths for NET-seq library NA19238. **D** Histogram of mapped read lengths for NET-seq library generated from HEK cells and published in Mayer *et al* [112]). **E** Pearson's correlations for NET-seq library coverage in gene bodies. Coverage calculated as number of reads mapping to gene standardized by gene length.

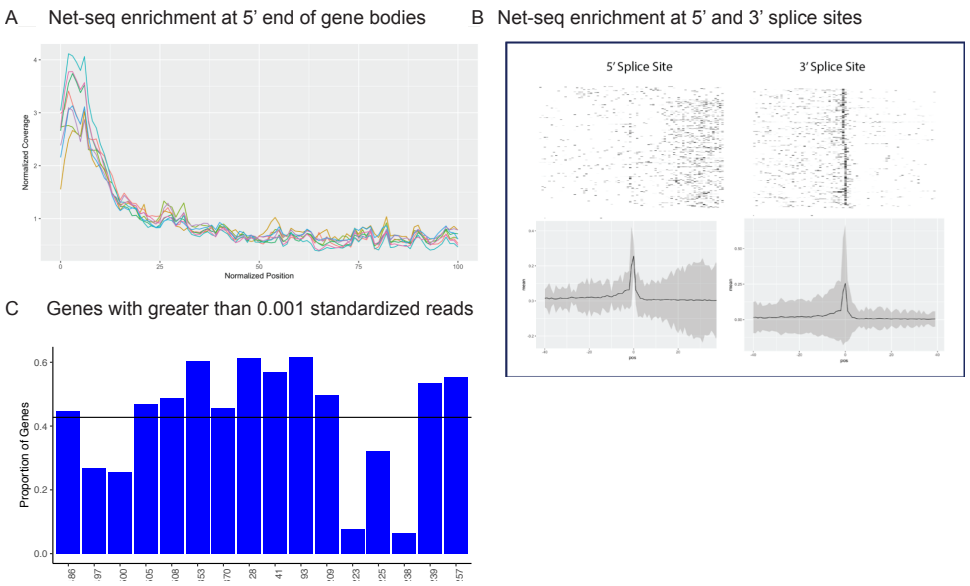


Figure 4.3: **NET-seq Gene coverage.** **A** Enrichment for Distribution of read coverage along gene bodies calculated with Picard tools (NA18505, NA18508, NA18486, NA19239, NA19239, NA19141, NA19193, NA19257, NA19128). **B** Histogram and smoothed density plots for NET-seq read coverage at 5' and 3' splice sites. **C** Proportion of genes in each NET-seq library with greater than 0.001 standardized reads (Methods).

genetic variation associated with variation in Pol II elongation rate. In turn, we next explored coverage at individual gene loci. We used a wavelet-based Empirical Bayes shrinkage method implemented in smashr to denoise genic signal ([197], Methods). For *ACTB*, the smoothing allowed us to identify regions of likely Pol II pausing at the transcription start site *(TSS)* and splice sites (Figure 4.4).

Smoothing to differentiate signal from random noise would not account for contamination by mature mRNAs or technical mapping errors. We found evidence of many genomic locations, such as in the *INSIG2* gene, with heavy read buildup. We were unable to identify technical or biological reasons for the high density of NET-seq reads (Figure 4.5). We hypothesize that unannotated chromatin associated mRNAs or low complexity repetitive reads. The protocol includes depletion of chromatin associated mature mRNAs, however the oligo pool is likely incomplete. Unannotated snoRNAs or snRNAs likely contribute to high density regions. Alternatively, because mapped reads are relatively short, repetitive genomic

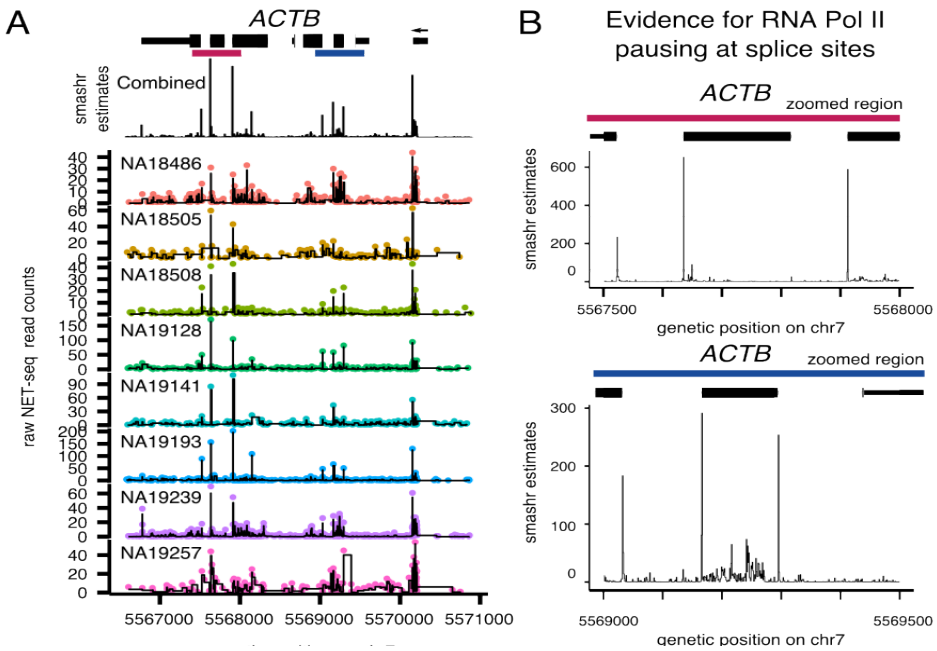


Figure 4.4: **Smoothing of NET-seq data using smashr [197].** **A** Raw counts along ACTB gene locus for 8 combined libraries (NA18505, NA18508, NA18486, NA19239, NA19239, NA19141, NA19193, NA19257, NA19128) as well as each library separately. **B** Smoothed coverage along *ACTB* on combined NET-seq data. Signal at 5' and 3' splice sites.

regions may be mipmapping, therefore contributing to regions of high read density.

We did not continue the analysis beyond these quality control metrics and low-level analyses. Further optimization of this protocol or another protocol is likely necessary to quantify Pol II pause pattern variation genome wide.

4.4 Discussion

By extracting and sequencing nascent chromatin-associated mRNA, we attempted to measure polymerase II (Pol II) density and estimate transcription elongation rates a population of human LCLs. Using data from 16 individuals we were able to capture the broad patterns previously described by Mayer *et al.* and others [112]. We found evidence for Poll II pausing at the TSS and at the 5' and 3' splice sites for highly expressed exons. We showed that in genes with high coverage, an Empirical Bayes shrinkage method could differentiate between

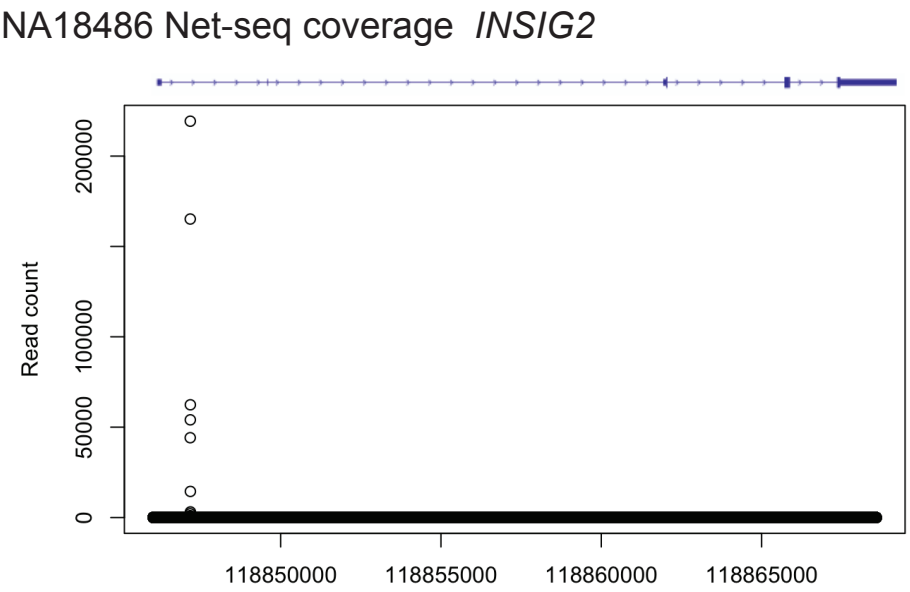


Figure 4.5: **NA18486 NET-seq coverage along INSIG2 locus** Read count (y axis) plotted by base pair coordinate on chromosome 2.

regions of high Pol II density and background random noise. Unfortunately, for most genes, we did not have high enough coverage to smooth the data. Further, we identified regions of the genome where a large proportion of the reads mapped for unknown technical reasons or biological contamination of mature chromatin associated mRNAs.

We believe the NET-seq libraries were of low quality and complexity for a number of reasons. The NET-seq protocol was difficult to optimize due to low concentration of input mRNA. Specifically, we needed to extract chromatin associated mRNA from 8 collections of 15 million cells to achieve the required $1\mu g$ of input RNA. Moreover, the protocol included multiple gel extraction steps where a large proportion of the mRNA was lost. While we explored alternative options, such as size selection with columns, none were specific enough for the desired fragments. Even after optimizing the protocol to achieve libraries of high enough concentration for sequencing, the nascent RNA fragments were shorter than reads published in Mayer *et al.* [112]. We believe that while we selected the optimal fragment lengths, shorter fragments were preferentially incorporated and amplified into libraries.

It is difficult to determine if our libraries were of lower quality than those published in

Mayer *et al.* because the published libraries also had a large number of un-mapped reads and a low signal-to-noise ratio. The authors largely characterized patterns across multiple genes and did not collect a population sample to identify variation. It is likely that sequencing coverage played a large role in the differences between our results and those reported in Mayer *et al.* [112]. For one replicated of HEK293T, they sequenced 1.2 billion reads with 555 million uniquely mapping. At this coverage, only 50% of the reads had coverage of over 1 read per kilobase per million *(RPKM)* [112]. While we acknowledged that sequencing to this depth would be unreasonable in a population sample, we planned to take advantage of shared genotypes to identify genetic variants associated with Pol II patterns. Unfortunately, our low coverage libraries were too noisy to confidently quantify Pol II density, even if we merged on genotypes.

We still believe that mapping genetic variants associated with Pol II density would likely help in our understanding on gene regulation. There are a variety of alternative methods to collect nascent mRNA or Pol II associated transcripts. Other potential methods largely fall into two classes. The first category relies on immunoprecipitation of Pol II or a specific post translationally modified version of Pol II [127, 56, 23, 81, 195, 113]. These methods potentially suffer from non-specific binding of anti-bodies, and with the exception of mNetseq, are restricted to 200bp resolution [127]. The second class of approaches utilize in-vitro incorporation of labeled nucleotides to identify regions of transcription potential. Transcription run-on assays, such as GRO-seq and PRO-seq, are vulnerable to variation in experimental conditions because the protocols include stopping and restarting transcriptions in non-physiological conditions [87, 37, 183, 55, 106, 113, 195].

To date, only one preprint has reported the usage of a variant of one of these methods to characterize nascent RNA in a population of human cell lines. *Kristjànsdòttir et al.*, quantified 5' capped nascent transcripts in 67 YRI LCLs with PRO-cap [85]. In contrast to our work, these authors aimed to characterize transcription at enhancers and to identify

genetic variation associated with enhancer transcription initiation. The authors also collected PRO-seq for 10 unique individuals [85]. It is possible that once published, we or others could use these data to test for genetic variation associated with co-transcriptional Pol II density variation.

4.5 Methods

4.5.1 Cell culture of LCLs

We cultured 16 human Epstein-Bar virus transformed lymphoblastoid cell lines (LCLs) in glutamine depleted RPMI (RPMI 1640 1X from Corning (15-040-CM)), completed with 15% FBS, 2mM GlutaMax (Gibco (35050-061)), 100 IU/ml Penicillin, and 100 $\mu g/mL$ Streptomyocin. We cultured all cells at 37C at 5% CO2. The 16 cell lines represent a subset of the YRI individual LCLs collected as part of the hapmap project and are available through Coriell [70]. We used lines NA19527, NA19239, NA19238, NA19225, NA19223, NA19209, NA19193, NA19128, NA19141, NA19128, NA18853, NA18508, NA18505, NA18501, NA18497, NA18486. These lines also represent a subset of those used for the 3' sequencing published in Chapter 2 and in Mittleman *et al.* [118].

4.5.2 Collections and library preparation

After growing the LCLs to around 1 million cells per ml, we separated $1.5x10^7$ cells into 10 tubes, one for total cells, one for nuclear fraction and eight for the chromatin fraction. Collection dates and details can be found in Mittleman *et al.* Additional File 2 [118]. We used the Native Elongating Transcript sequencing (NET-seq) collection protocol published in Mayer and Churchman 2016 with minor adjustments to three buffers [114]. Specifically, we added 1M $MgCl_2$ and 100% Glycerol to the Cytoplasmic lysis buffer, Sucrose buffer, and the Nuclei wash buffer. We added the $MgCl_2$ to stabilize the nucleus and the glycerol as

a freezing protectant. We halted transcription with α-amanatin and separated the nuclear fraction using mild detergent and a sucrose cushion. We then separate the nucleoplasm from the chromatin using urea, salt, and a mild detergent. We then collected the chromatin through centrifugation and degradation of the DNA with a DNase treatment. We used the Qiagen miRNAeasy kit with manufacture instructions to extract mRNA from all three fractions.

We generated NET-seq library according to the Mayer and Churchman protocol with custom oligos ordered from IDT [114]. I captured the 3' end of chromatin associated mRNA molecules with a barcoded linker and convert the fragments into cDNA for library preparation. We sequenced each NET-seq library at the University Genomics Core facility using single end 50bp sequencing on the Illumina HiSeq4000 machine. We multiplexed 8 libraries together and sequenced each group on a total of 3 lanes. Custom sequencing primers can be found in Mayer and Churchman protocol [114].

4.5.3 Data processing

We mapped all NET-seq libraries to GRCH37.75 downloaded from Ensemble using subjunc with default settings [71, 97] We used umi_tools extract to extract the 6 base UMIs and umi-tools dedup to collapse duplicate reads [169].

We assessed computed genome coverage at basepair resolution using bedtools genomecov with the -d and -5 flags. We measured the coverage density along the gencode.v19 gene annotation using picard CollectRnaSeqMetrics [52, 152]. We used featureCounts with the -T 5 flag to quantify reads within the gencode.v19 gene annotation [52, 99]. We used pysam to extract mapped read statistics from bam files [91]. We downloaded the HEK NET-seq published in Mayer *et al.* available from GEO under accession number GSE61332 [111]. We re-processed the fastq file using our mapping pipeline. We used the smash.pois function with the EM algorithm in the smashr package on individual genes for signal denoising [197]. The

function implemented a wavelet-based Empirical Bayes shrinkage method [197].

We attempted to recreate a version of Mayer *et al.* figure 7 using NA18486 sequence coverage for this analysis [111]. Due to low library coverage and complexity, we did not separate the exons into, constitutive, alternatively retained, alternative skipped. We quantified coverage at base pair resolution for 40bp upstream and downstream of the 5' and 3' splice sites of the top 5% covered exons. We then standardized base pair coverage by exon coverage.

CHAPTER 5

CONCLUSION

Quantitative genetics and comparative primate functional genomic approaches can both be used to disentangle the molecular mechanisms regulating gene expression. Firstly, by identifying genetic variation correlated with a range of molecular functions, quantitative trait loci *(QTL)* mapping studies have contributed to models of how variation percolates through the gene regulatory cascade [42, 116, 95, 144]. Secondly, to establish how gene regulatory mechanisms contribute to species divergence, regulatory features have been characterized and compared in humans and non-human primates [137, 159, 49, 15, 80, 135, 166] Both approaches have highlighted the importance of the chromatin accessibility state of cis-regulatory elements, such as promoters and enhancers, in controlling gene expression. However, with the exception of some work on alternative splicing, both lines of research have left co-transcriptional gene regulatory mechanisms that act through variation in mRNA isoforms largely understudied [17, 95].

Most human genes have the potential to express isoforms terminating at alternative polyadenylation *(APA)* sites with distinct downstream regulatory fates. Usage of an alternative polyadenylation sites *(PAS)* in the 3' UTR can lead to differential inclusion of cis-regulatory elements, such as miRNA binding sites and RNA binding protein motifs. A transcript terminating at an intronic PAS may be subject to decay or be translated into a functional protein. Both 3' UTR and intronic APA can cause isoform-specific mRNA stability, mRNA localization, and translation efficiency. (reviewed [179]) Given its downstream consequences, APA likely plays a significant role in gene regulation.

The complex ways genetic variation can act through APA to shape the transcriptome and proteome have not previously been described. Likewise, how functional divergence in APA between human and chimpanzee genomes contributes to differential expression of genes has not been explored. Thus, to address this gap, in this dissertation, I have explored the role

of APA in contributing to transcriptome and proteome diversity in a population of human cell lines and in a panel of human and chimpanzee cell lines.

In Chapter 2, I used a quantitative trait locus *(QTL)* mapping approach to identify genetic variation associated with APA. By measuring QTL sharing between APA and other molecular phenotypes, I determined that genetic variation likely acts through APA to regulate mRNA expression, translation, and protein levels (Figure 5.1-Top). In Chapter 3, I took a comparative primate genomics approach to understand the functional conservation of APA as a regulatory mechanism. Through this analysis, I also gained the power to dissect the details of how APA interacts with other regulatory mechanisms at a finer scale than in Chapter 2 (Figure 5.1-Bottom). While the molecular signals for APA have been characterized, variation in PAS signal sites cannot explain APA variation genome-wide. We hypothesized that co-transcriptional mechanisms, such as RNA polymerase II *(Pol II)* elongation rate, could explain APA variation. In Chapter 4, I tried to formally establish this link in the gene regulatory cascade by identifying genetic variation association with Pol II pausing.

5.1 Genetic underpinning of APA variation

In Chapter 2, I worked under the supervision of Dr. Yang Li to study the impact of genetic variation on APA. I collected 3' sequencing *(3' Seq)* data on mRNA extracted from whole cells and nuclei of 52 human lymphoblastoid cell lines *(LCLs)*. Using these data, I was able to show that genetic effects on APA largely act on PAS choice. Because we have previously identified genetic variants associated with mRNA expression, translation, and protein levels in the same population, I was able to integrate apaQTLs, eQTLs, riboQTLs, and pQTLs into my analysis and give a comprehensive overview of how genetic variation acts through APA to mediate gene expression differencs. I showed that alleles associated with increased usage of intronic PAS also correlated with decreased mRNA expression. My work suggested that, through usage of intronic PAS and other mechanisms, APA can explain around 20% of the

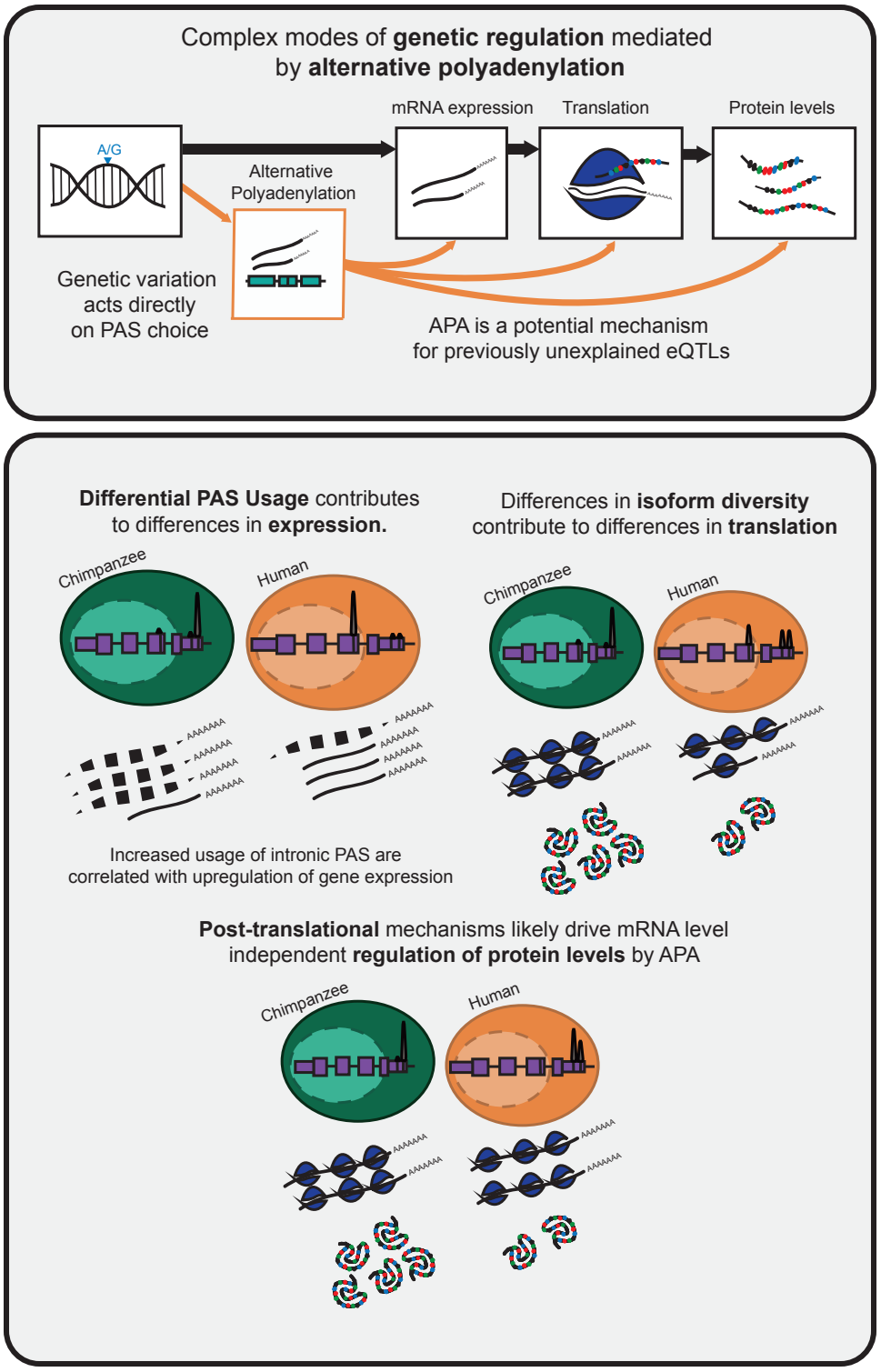


Figure 5.1: Graphical abstract to demonstrate major findings in this dissertation. Top: Modes by which genetic variation can influence gene regulation through APA. Bottom: Differences in APA between human and chimpanzee help explain differences in mRNA expression, translation, and protein levels between species.

eQTLs previously unexplained by chromatin variation [95]. I also identified apaQTLs that were not eQTLs. Some of the apaQTLs that are not significantly associated with expression are correlated with variation in translation, protein levels, or both. Finally, I explored the link between APA and complex traits. I demonstrated that genetic variation surrounding PAS significantly contributes to the heritability of complex immune traits and that 19.3% of apaQTLs are in high linkage disequilibrium with GWAS loci.

Throughout my thesis work, a number of groups have recognized the opportunity to study the genetic variation associated with APA [199, 108, 92]. My work represents a novel contribution to the field in that I directly measured PAS usage with 3' Seq rather than estimating usage from traditional RNA sequencing data. While PAS can be inferred from traditional RNA sequencing methods, estimates of usage are imprecise [62]. To our knowledge, we are the only group to measure PAS usage in both nuclear and total mRNA and to conclude that genetic variation largely acts on PAS choice rather than through isoform specific regulation. By detecting apaQTLs in a range of tissues and cancer cell lines, these studies further implicated the role of genetic variation acting through APA to contribute to complex trait variation and disease risk [199, 108, 92].

Alternative mRNA splicing and APA are both co-transcriptional gene regulatory mechanisms. Interplay between these two processes contribute to gene regulatory outcomes because splicing and polyadenylation rely on a subset of similar protein complexes competing for space [150]. Although I chose not to explore the interplay between mRNA splicing and APA, I used my data to test one interesting hypothesis. Previous studies have implicated the splice factor, U1 snRNP, in protecting introns from premature cleavage and polyadenylation, termed telescripting [76, 13, 128]. Because the U1 snRNP binds to 5' splice sites, I expected that some intronic PAS to escape telescripting because they lie in introns with weak 5' splice sites. Supporting this hypothesis, I found an enrichment of intronic PAS in introns with the weakest 5' splice sites. However, if and how genetic variation can contribute to both

alternative splicing and APA remain to be known.

5.2 Functional conservation of APA

In Chapter 2, I showed that some genetic variation associated with APA is also associated with mRNA expression variation. In Chapter 3, I used a comparative genomics approach to further investigate the role of APA in gene regulation. I concluded that differential expression could be driven specifically by switching between dominant isoforms. However, I did not find evidence that differences isoform diversity contributes to differential expression. I also provided additional evidence that APA contributes to proteomic diversity independent from differences in mRNA levels.

In Chapter 2, I showed a negative correlation between intronic apaQTL and eQTL effect sizes. Specifically, for several loci, the allele associated with increased intronic PAS usage was also associated with increased mRNA expression. When I examined a similar relationship between human and chimpanzee, I detected a positive correlation between usage of intronic PAS and gene expression. In Chapter 3's discussion, I outlined a set of non-mutually exclusive interpretations for these seemingly inconsistent results. First, increased usage of an intronic PAS could lead to rapid nonstop decay the transcript. Second, usage of an intronic PAS could allow the transcript to evade down regulation through cis-regulatory motifs located in the 3' UTR. While I cannot say for certain which mechanism is more active genome wide, the large effect sizes for 3' UTR differences between species have led me to place more weight on the second interpretation.

With this in mind, functional analysis of genes with intronic and 3' UTR PAS are necessary to better understand the mechanistic relationship between intronic PAS and mRNA expression. However, functional follow up experiments may be complicated because each PAS usage ratio assigned to the same gene is not an independent measurement. Without independent measurements for each site, it is difficult to identify which isoform contributes

most to variation in mRNA expression. In other words, it is possible that decreased usage of the 3' UTR sites correlated with decreased mRNA expression, but we detected the intronic relationship as more significant.

5.3 Pol II dynamics to understand co-transcriptional gene regulation

In Chapter 4, I described a project that we unfortunately could not complete. We optimized native elongation transcript sequencing *(Net-seq)* for human LCLs with the goal of mapping genetic variation to variation in Pol II elongation rate. We concluded that the technology was not advanced enough to answer the questions we intended to address. Nonetheless, by optimizing Net-seq, I validated a protocol to extract mRNA from nuclei. In addition, I also demonstrated the usefulness of the wavelet-based Empirical Bayes shrinkage method to de-noise functional genomic data [197].

Had we been able to quantify Pol II density variation with Net-seq, it would have informed analyses for my other projects. For example, in Chapter 2, I identified genetic variation associated with APA but did not ask more detailed questions about how the variants control APA. Given the evidence that apaQTLs are enriched in regions of elongation and around the PAS that they regulate, I hypothesize that genetic variation associated with Pol II elongation rate are also apaQTLs. If my results had supported this hypothesis, the results would have added to the growing literature about transcription kinetics and APA.

Transcription kinetics have previously been implicated in APA (reviewed in [149]). Interestingly, experiments testing for the causal relationship between Pol II elongation rate and PAS usage are inconsistent. For example, Pol II mutations leading to a slowdown in elongation resulted in proximal PAS usage, suggesting changes in Pol II elongation speed cause a PAS to be used [146, 84, 41, 68]. Conversely, nascent transcription measurements suggested that cleavage and polyadenylation factors may be responsible for Pol II pausing

rather than the other way around. Specifically, Nojima *et al.* measured the density of post-transcriptionally modified Pol lI after the depletion of cleavage and polyadenylation factors, CPSF73 and CstF64. In each knockdown experiment, the authors reported a reduction of Pol II density at transcription end sites. They concluded that Pol II pausing at transcription end sites is dependent on the polyadenylation machinery [127]. Although I may have been able to identify the general sharing between genetic effects on Pol II elongation rate and APA, additional follow-up would have been necessary to prove causality.

Recent investigations have implicated DNA methylation in the relationship between Pol II elongation and APA control [126]. Nanavaty *et al.*, explored the role of DNA methylation and APA by measuring PAS usage in cell lines with normal and heavily depleted methylation profiles. The authors annotated genes whereby methylation of CTCF binding domains directly downstream of proximal PAS prevented usage of proximal PAS in favor of downstream sites. Interestingly, in the demethylated state, the authors reported increased Pol II density at the CTCF region [126]. The authors concluded that methylation of DNA may contribute to both transcription elongation and APA regulation. Because the authors identified a range of PAS usage changes upon demethylation of PAS regions, their proposed mechanism does not explain APA genome wide. Exploring the relationship between genetic variation associated with DNA methylation levels and apaQTLs may point to additional mechanisms contributing to co-transcriptional transcriptional gene regulation.

5.4 Future directions

5.4.1 Expanding work to new biological processes

I collected the data presented in Chapters 2 - 4 from LCLs growing in culture at a steady state. This work represents the first step to understanding the gene regulatory effects of APA, but we need to expand this work further to more cell types and conditions. The

methods and analysis pipelines presented here provide a framework for future studies to explore both the genetic underpinnings and evolutionary trajectory of APA.

A dynamic analysis of how genetic variation impacts APA would add another dimension to our understanding of gene regulatory pathways. As global changes in APA have been characterized for a range of dynamic biological processes, it is possible that genetic effects acting on such processes would be distinct from those identified in steady state. Changes in cellular state correspond with global changes in APA regulation. Ji and Tian detected global 3' UTR shortening upon reprogramming of human and mouse somatic cells to stem cells [72]. Consistent with this observation, a number of groups have observed 3' UTR lengthening during development of a range of species [72, 64, 93, 124]. While changes in expression levels of cleavage and polyadenylation factors likely contribute to APA during cellular state changes, the genetic control of these processes is unknown.

Using the panel of induced pluripotent stem cells *iPSCs* established and validated in the Gilad lab, we could identify genetic variation associated with dynamic changes in alternative polyadenylation during cellular differentiation. As a proof of principle, Strober, Elorbany and Rhodes *et al.* detected transient and dynamic eQTLs during differentiation of iPSCs into cardiomyocytes [172]. These results suggest that transient genetic effects also contribute to complex traits and disease. Given the work outlined in this disserations, I would expect that characterization of APA along a similar time course would contribute to our understanding of the connection between APA, other gene regulatory process, and complex traits.

In Chapter 3, I showed that APA is largely conserved between human and chimpanzee. Yet, the divergent PAS that I identified helped to inform differences in mRNA expression, translation, and protein levels across species. Given that APA contributes to gene regulatory changes in response to stress, it would be informative to investigate whether APA is more or less conserved between species in stressful conditions. Also, because humans and chimpanzees have similar heart anatomy but different cardiovascular disease *(CVD)* pathol-

ogy, it would be interesting to ask if differences in APA upon the stresses associated with CVD contribute to pathological differences [88, 186]. Ward and Gilad subjected human and chimpanzee iPSC derived cardiomyocytes to a hypoxia and re-oxygenation protocol to ask if differences in mRNA expression contribute to CVD phenotypic differences. Of the 32 cleavage and polyadenylation factors, Ward and Gilad measured expression of 20 [160, 193]. Of the 20 genes, 7 showed a conserved response and 1 showed a chimpanzee-specific response. If expression of these polyadenylation factors changes in response to hypoxia and reoxygenation, I also expect both changes in conserved and species-specific APA. By studying APA conservation or divergence in a similar framework as Ward and Gilad, we may uncover additional gene regulatory mechanisms contributing to differences in CVD pathology.

3' Seq and apaQTL calling in cancer tissues or cell lines to identify genetic variation associated with intronic PAS usage is an important future direction for my work. In contrast to APA variation associated with 3' UTR length differences, intronic polyadenylation leads to coding changes that may or may not result in functional proteins [179, 187, 202]. In Chapters 2 and 3, I described work characterizing the downstream regulatory consequences of intronic polyadenylation. In 2018, Lee *et al.* presented a mechanism for cancer pathogenesis through intronic polyadenylation that leads to partial or full inactivation of tumor suppressor genes [90]. In Chapter 2, I identified an apaQTL in the ELL2 gene *(rs56219066)* associated with use of an intronic PAS that has previously been associated with risk for multiple myeloma [173]. It is likely that expanding my work to cancer cell lines would lead to additional insight into the ways genetic variation can act through APA to modify cancer risk.

Using available RNA sequencing data in cancer cell lines and tissues, apaQTLs associated with 3' UTR length have been identified. Earlier this year, Yang *et al.* released the SNP2APA database with 467,942 cis-apaQTLs and 30,721 trans-apaQTLs identified in 32 cancer types [200]. Rather than calculating usage of each PAS, the authors tested for genetic variation associated with the percentage of distal polyA site usage index *(PDUI)*. This strategy likely

identified apaQTLs associated with 3' UTR length and is a necessary first step toward understanding the role of APA regulation and cancer.

5.4.2 Technical and methodological future directions

In this dissertation, I relied on correlations and enrichments to draw conclusions of the role of APA in gene regulation. In Chapter 2, I identified genetic variants correlated with variation in APA. I then asked if the same variants are correlated with mRNA expression, translation, and protein variation. Using this strategy, I concluded that genetic variation acts through APA to regulate gene expression. If the work in Chapter 4 came to fruition, I would have used a similar approach to relate Pol II elongation rate to differences in APA and splicing. In Chapter 3, I suggested APA may be driving differences in gene regulation between species based on enrichments of statistically significant differences between species in each of the tested molecular phenotypes.

While correlations and enrichments help us gain important insight, I was not able to report causation or the percent of variation explained by APA in other regulatory phenotypes. For example, regulatory QTL and comparative genomic studies, mediation analyses have been used to test if genetic variation causes changes in a molecular mechanism which then causes differences in gene expression [145, 141, 15, 49]. In my context, I would have asked if genetic variation caused changes in APA and if the changes in APA caused changes gene expression. We have methods for mediation analysis when effect sizes for the mediator and outcome are measured on the same scale. For example, DNA methylation and mRNA expression are both measured linearly, as a sum of genomic reads. However, APA is a ratio of reads and is thus on a different scale than gene expression phenotypes, including mRNA and protein expression. New mathematical models are necessary to incorporate effect sizes measured as ratios into traditional mediation frameworks.

Long read mRNA sequencing with Pacific Biosciences *(PacBio)* SMRT technology or

Oxford Nanopores Ion technology presents a great opportunity for studying APA [3]. By capturing and sequencing full transcripts, we would be able to detect PAS with higher confidence. In addition, by sequencing full transcripts, we would also be able to connect APA to alternative splicing [3, 156]. However, long read sequencing has a higher per-base error rate and may be subject to mapping errors due to template-switching artifacts [7]. Thus, it is likely that combining long and short read sequencing may provide more confidence for PAS detection and quantification. As a small part of my dissertation work, I submitted a subset of the human LCLs that I used in Chapter 2 for PacBio Iso-seq sequencing. While these data have yet to be fully analyzed, I hope that future work will combine the data published with this dissertation with the PacBio dataset to develop best practices for PAS detection and quantification.

5.5 Concluding remarks

Exploration of gene regulation at an isoform level adds an additional layer of complexity to our understanding of gene expression. While many groups have started to study the regulatory causes and effects of alternative splicing, less has been done to understand APA genome wide. This dissertation and the resulting publications are an attempt to begin filling this gap. I identified genetic variation associated with APA in order to test if APA differences contribute to differences in mRNA expression, translation, and protein levels. I then took a comparative genomic approach to further disentangle the relationship between APA and gene regulation. In doing so, I helped to fill in some of the many remaining gaps in our ability to read the gene regulatory code as a way to understand human complex traits and diseases.

References

[1] Karen Adelman and John T. Lis. Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans. *Nature Reviews. Genetics*, 13(10):720–731, 2012.

[2] François Aguet, Alvaro N. Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E. Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, Yanyu Liang, Meritxell Oliva, Princy E. Parsana, Elise Flynn, Laure Fresard, Eric R. Gaamzon, Andrew R. Hamel, Yuan He, Farhad Hormozdiari, Pejman Mohammadi, Manuel Muñoz-Aguirre, YoSon Park, Ashis Saha, Ayellet V. Segrć, Benjamin J. Strober, Xiaoquan Wen, Valentin Wucher, Sayantan Das, Diego Garrido-Martín, Nicole R. Gay, Robert E. Handsaker, Paul J. Hoffman, Seva Kashin, Alan Kwong, Xiao Li, Daniel MacArthur, John M. Rouhana, Matthew Stephens, Ellen Todres, Ana Viñuela, Gao Wang, Yuxin Zou, The GTEx Consortium, Christopher D. Brown, Nancy Cox, Emmanouil Dermitzakis, Barbara E. Engelhardt, Gad Getz, Roderic Guigo, Stephen B. Montgomery, Barbara E. Stranger, Hae Kyung Im, Alexis Battle, Kristin G. Ardlie, and Tuuli Lappalainen. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, page 787903, 2019.

[3] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):30, 2020.

[4] 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.

[5] William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H. A. Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo-de-Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, and Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5):1415–1429.e19, 2016.

[6] Jan Attig, Igor Ruiz de los Mozos, Nejc Haberman, Zhen Wang, Warren Emmett, Kathi Zarnack, Julian König, and Jernej Ule. Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *eLife*, 5:e19545, 2016.

[7] Zsolt Balázs, Dóra Tombácz, Zsolt Csabai, Norbert Moldován, Michael Snyder, and Zsolt Boldogkői. Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics*, 20(1):824, 2019.

[8] Nicholas E. Banovich, Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F. Degner, John D. Blischak, Julien Roux, Jonathan K. Pritchard, and Yoav Gilad. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genetics*, 10(9):e1004663, 2014.

[9] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

[10] Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Genomic variation. Impact of regulatory variation from RNA to protein. *Science (New York, N.Y.)*, 347(6222):664–667, 2015.

[11] Alexis Battle, Zia Khan, Sidney H. Wang, Amy Mitrano, Michael J. Ford, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, 2015.

[12] Emmanuel Beaudoing, Susan Freier, Jacqueline R. Wyatt, Jean-Michel Claverie, and Daniel Gautheret. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research*, 10(7):1001–1010, 2000.

[13] Michael G. Berg, Larry N. Singh, Ihab Younis, Qiang Liu, Anna Maria Pinto, Daisuke Kaida, Zhenxi Zhang, Sungchan Cho, Scott Sherrill-Mix, Lili Wan, and Gideon Dreyfuss. U1 snRNP Determines mRNA Length and Regulates Isoform Expression. *Cell*, 150(1):53–64, 2012.

[14] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James A Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–1048, 2010.

[15] Lauren E. Blake, Julien Roux, Irene Hernando-Herraez, Nicholas E. Banovich, Raquel Garcia Perez, Chiaowen Joyce Hsiao, Ittai Eres, Claudia Cuevas, Tomas Marques-Bonet, and Yoav Gilad. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Research*, 30(2):250–262, 2020.

[16] M. Blanchette. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14(4):708–715, 2004.

[17] Ran Blekhman, John C. Marioni, Paul Zumbo, Matthew Stephens, and Yoav Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Research*, 20(2):180–189, 2010.

[18] Ran Blekhman, Alicia Oshlack, Adrien E. Chabot, Gordon K. Smyth, and Yoav Gilad. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS genetics*, 4(11):e1000271, 2008.

[19] Ran Blekhman, Alicia Oshlack, and Yoav Gilad. Segmental Duplications Contribute to Gene Expression Differences Between Humans and Chimpanzees. *Genetics*, 182(2):627–630, 2009.

[20] Marc Jan Bonder, Craig Smail, Michael J. Gloudemans, Laure Frésard, David Jakubosky, Matteo D'Antonio, Xin Li, Nicole M. Ferraro, Ivan Carcamo-Orive, Bogdan Mirauta, Daniel D. Seaton, Na Cai, Danilo Horta, HipSci Consortium, iPSCORE Consortium, GENESiPS Consortium, PhLiPS Consortium, Erin N. Smith, Kelly A. Frazer, Stephen B. Montgomery, and Oliver Stegle. Systematic assessment of regulatory effects of human disease variants in pluripotent cells. *bioRxiv*, page 784967, 2019.

[21] Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome Institute at Washington University, Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke, Brian J. Raney, Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[22] Christopher Buccitelli and Matthias Selbach. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 2020.

[23] Michael J. Buck and Jason D. Lieb. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, 2004.

[24] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109:21.29.1–21.29.9, 2015.

[25] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

[26] Carolyn E. Cain, Ran Blekhman, John C. Marioni, and Yoav Gilad. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics*, 187(4):1225–1234, 2011.

[27] Minal Çalışkan, Darren A. Cusanovich, Carole Ober, and Yoav Gilad. The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics*, 20(8):1643–1652, 2011.

[28] Minal Çalışkan, Jonathan K. Pritchard, Carole Ober, and Yoav Gilad. The Effect of Freeze-Thaw Cycles on Gene Expression Levels in Lymphoblastoid Cell Lines. *PLoS ONE*, 9(9):e107166, 2014.

[29] Fernando Carrillo Oesterreich, Stephan Preibisch, and Karla M. Neugebauer. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Molecular Cell*, 40(4):571–581, 2010.

[30] Sean B. Carroll. Evolution at two levels: On genes and form. *PLoS biology*, 3(7):e245, 2005.

[31] David Carter, Lyubomira Chakalova, Cameron S. Osborne, Yan-feng Dai, and Peter Fraser. Long-range chromatin regulatory interactions in vivo. *Nature Genetics*, 32(4):623–626, 2002.

[32] Joel M. Chick, Steven C. Munger, Petr Simecek, Edward L. Huttlin, Kwangbom Choi, Daniel M. Gatti, Narayanan Raghupathy, Karen L. Svenson, Gary A. Churchill, and Steven P. Gygi. Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, 534(7608):500–505, 2016.

[33] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[34] 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.

[35] Deanna M. Church, Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M. McLaren, Graham R. S. Ritchie, Derek Albracht, Milinn Kremitzki, Susan Rock, Holland Kotkiewicz, Colin Kremitzki, Aye Wollam, Lee Trani, Lucinda Fulton, Robert Fulton, Lucy Matthews, Siobhan Whitehead, Will Chow, James Torrance, Matthew Dunn, Glenn Harden, Glen Threadgold, Jonathan Wood, Joanna Collins, Paul Heath, Guy Griffiths, Sarah Pelan, Darren Grafham, Evan E. Eichler, George Weinstock, Elaine R. Mardis, Richard K. Wilson, Kerstin Howe, Paul Flicek, and Tim Hubbard. Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091, 2011.

[36] D. F. Colgan and J. L. Manley. Mechanism and regulation of mRNA polyadenylation. *Genes & Development*, 11(21):2755–2766, 1997.

[37] Leighton J. Core, André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, 46(12):1311–1320, 2014.

[38] Gábor Csárdi, Alexander Franks, David S. Choi, Edoardo M. Airoldi, and D. Allan Drummond. Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *PLOS Genetics*, 11(5):e1005206, 2015.

[39] Daniel S. Day, Bing Zhang, Sean M. Stevens, Francesco Ferrari, Erica N. Larschan, Peter J. Park, and William T. Pu. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biology*, 17(1):120, 2016.

[40] P de Bie and A Ciechanover. Ubiquitination of E3 ligases: Self-regulation of the ubiquitin system via proteolytic and non-proteolytic mechanisms. *Cell Death & Differentiation*, 18(9):1393–1402, 2011.

[41] Manuel de la Mata, Claudio R Alonso, Sebastián Kadener, Juan P Fededa, Matías Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R Kornblihtt. A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Molecular Cell*, 12(2):525–532, 2003.

[42] 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, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[43] 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.

[44] Dafne Campigli Di Giammartino, Kensei Nishida, and James L. Manley. Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell*, 43(6):853–866, 2011.

[45] 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.

[46] Sara J. Dubbury, Paul L. Boutz, and Phillip A. Sharp. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, 564(7734):141–145, 2018.

[47] Tatyana Dubnikov, Tziona Ben-Gedalya, and Ehud Cohen. Protein Quality Control in Health and Disease. *Cold Spring Harbor Perspectives in Biology*, 9(3):a023523, 2017.

[48] Wolfgang Enard, Anne Fassbender, Fabian Model, Péter Adorján, Svante Pääbo, and Alexander Olek. Differences in DNA methylation patterns between humans and chimpanzees. *Current biology: CB*, 14(4):R148–149, 2004.

[49] Ittai E. Eres, Kaixuan Luo, Chiaowen Joyce Hsiao, Lauren E. Blake, and Yoav Gilad. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genetics*, 15(7):e1008278, 2019.

[50] Benjamin J. Fair, Lauren E. Blake, Abhishek Sarkar, Bryan J. Pavlovic, Claudia Cuevas, and Yoav Gilad. Gene expression variability in human and chimpanzee populations share common determinants. Technical report, 2020.

[51] Stephen N Floor and Jennifer A Doudna. Tunable protein synthesis by transcript isoforms in human cells. *eLife*, 5:e10921, 2016.

[52] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T. Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G. Izuogu, Julien Lagarde, Fergal J. Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C. P. Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M. Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S. Choudhary, Mark Gerstein, Roderic Guigó, Tim J. P. Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L. Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2019.

[53] Becky Fusby, Soojin Kim, Benjamin Erickson, Hyunmin Kim, Martha L. Peterson, and David L. Bentley. Coordination of RNA Polymerase II Pausing and 3' End Processing Factor Recruitment with Alternative Polyadenylation. *Molecular and Cellular Biology*, 36(2):295–303, 2016.

[54] Daniel J Gaffney, Jean-Baptiste Veyrieras, Jacob F Degner, Roger Pique-Regi, Athma A Pai, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, 13(1):R7, 2012.

[55] Alessandro Gardini. Global Run-On Sequencing (GRO-Seq). *Methods in Molecular Biology (Clifton, N.J.)*, 1468:111–120, 2017.

[56] P. Gariglio, M. Bellard, and P. Chambon. Clustering of RNA polymerase B molecules in the $5'$ moiety of the adult β-globin gene of hen erythrocytes. *Nucleic Acids Research*, 9(11):2589–2598, 1981.

[57] Yoav Gilad, Alicia Oshlack, Gordon K. Smyth, Terence P. Speed, and Kevin P. White. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, 440(7081):242–245, 2006.

[58] David Gokhman, Malka Nissim-Rafinia, Lily Agranat-Tamir, Genevieve Housman, Raquel García-Pérez, Esther Lizano, Olivia Cheronet, Swapan Mallick, Maria A. Nieves-Colón, Heng Li, Songül Alpaslan-Roodenberg, Mario Novak, Hongcang Gu, Jason M. Osinski, Manuel Ferrando-Bernal, Pere Gelabert, Iddi Lipende, Deus Mjungu, Ivanela Kondova, Ronald Bontrop, Ottmar Kullmer, Gerhard Weber, Tal Shahar, Mona Dvir-Ginzberg, Marina Faerman, Ellen E. Quillen, Alexander Meissner, Yonatan Lahav, Leonid Kandel, Meir Liebergall, María E. Prada, Julio M. Vidal, Richard M. Gronostajski, Anne C. Stone, Benjamin Yakir, Carles Lalueza-Fox, Ron Pinhasi, David Reich, Tomas Marques-Bonet, Eran Meshorer, and Liran Carmel. Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nature Communications*, 11(1):1189, 2020.

[59] R. R. Graham, C. Kyogoku, S. Sigurdsson, I. A. Vlasova, L. R. L. Davies, E. C. Baechler, R. M. Plenge, T. Koeuth, W. A. Ortmann, G. Hom, J. W. Bauer, C. Gillett, N. Burtt, D. S. Cunninghame Graham, R. Onofrio, M. Petri, I. Gunnarsson, E. Svenungsson, L. Ronnblom, G. Nordmark, P. K. Gregersen, K. Moser, P. M. Gaffney, L. A. Criswell, T. J. Vyse, A.-C. Syvanen, P. R. Bohjanen, M. J. Daly, T. W. Behrens, and D. Altshuler. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proceedings of the National Academy of Sciences*, 104(16):6758–6763, 2007.

[60] Natalia Gromak, Steven West, and Nick J. Proudfoot. Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Molecular and Cellular Biology*, 26(10):3986–3996, 2006.

[61] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[62] Kevin C. H. 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] 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.

[64] Valérie Hilgers, Michael W. Perry, David Hendrix, Alexander Stark, Michael Levine, and Benjamin Haley. Neural-specific elongation of 3' UTRs during Drosophila development. *Proceedings of the National Academy of Sciences of the United States of America*, 108(38):15864–15869, 2011.

[65] Peter V. Hornbeck, Bin Zhang, Beth Murray, Jon M. Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(Database issue):D512–520, 2015.

[66] Genevieve Housman and Yoav Gilad. Prime time for primate functional genomics. *Current Opinion in Genetics & Development*, 62:1–7, 2020.

[67] Genevieve Housman, Lorena M. Havill, Ellen E. Quillen, Anthony G. Comuzzie, and Anne C. Stone. Assessment of DNA Methylation Patterns in the Bone and Cartilage of a Nonhuman Primate Model of Osteoarthritis. *CARTILAGE*, 10(3):335–345, 2019.

[68] K. J. Howe. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA*, 9(8):993–1006, 2003.

[69] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[70] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[71] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[72] Zhe Ji, Ju Youn Lee, Zhenhua Pan, Bingjun Jiang, and Bin Tian. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17):7028–7033, 2009.

[73] Roby Joehanes, Xiaoling Zhang, Tianxiao Huan, Chen Yao, Sai-xia Ying, Quang Tri Nguyen, Cumhur Yusuf Demirkale, Michael L. Feolo, Nataliya R. Sharopova, Anne Sturcke, Alejandro A. Schäffer, Nancy Heard-Costa, Han Chen, Po-ching Liu, Richard Wang, Kimberly A. Woodhouse, Kahraman Tanriverdi, Jane E. Freedman, Nalini Raghavachari, Josée Dupuis, Andrew D. Johnson, Christopher J. O'Donnell, Daniel Levy, and Peter J. Munson. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology*, 18(1):16, 2017.

[74] Iris Jonkers, Hojoong Kwak, and John T. Lis. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, 3:e02407, 2014.

[75] Goo Jun, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics*, 91(5):839–848, 2012.

[76] Daisuke Kaida, Michael G. Berg, Ihab Younis, Mumtaz Kasim, Larry N. Singh, Lili Wan, and Gideon Dreyfuss. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, 468(7324):664–668, 2010.

[77] Mazen W. Karaman, Marlys L. Houck, Leona G. Chemnick, Shailender Nagpal, Daniel Chawannakul, Dominick Sudano, Brian L. Pike, Vincent V. Ho, Oliver A. Ryder, and Joseph G. Hacia. Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. *Genome Research*, 13(7):1619–1630, 2003.

[78] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[79] Philipp Khaitovich, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. Evolution of primate gene expression. *Nature Reviews Genetics*, 7(9):693–702, 2006.

[80] Z. Khan, M. J. Ford, D. A. Cusanovich, A. Mitrano, J. K. Pritchard, and Y. Gilad. Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures. *Science*, 342(6162):1100–1104, 2013.

[81] Tae Hoon Kim, Leah O. Barrera, Ming Zheng, Chunxu Qu, Michael A. Singer, Todd A. Richmond, Yingnian Wu, Roland D. Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005.

[82] M C King and A C Wilson. Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, 188(4184):107–16, 1975.

[83] Derek Klarin, Scott M. Damrauer, Kelly Cho, Yan V. Sun, Tanya M. Teslovich, Jacqueline Honerlaw, David R. Gagnon, Scott L. DuVall, Jin Li, Gina M. Peloso, Mark Chaffin, Aeron M. Small, Jie Huang, Hua Tang, Julie A. Lynch, Yuk-Lam Ho, Dajiang J. Liu, Connor A. Emdin, Alexander H. Li, Jennifer E. Huffman, Jennifer S. Lee, Pradeep Natarajan, Rajiv Chowdhury, Danish Saleheen, Marijana Vujkovic, Aris Baras, Saiju Pyarajan, Emanuele Di Angelantonio, Benjamin M. Neale, Aliya Naheed, Amit V. Khera, John Danesh, Kyong-Mi Chang, Gonçalo Abecasis, Cristen Willer, Frederick E. Dewey, David J. Carey, Global Lipids Genetics Consortium, Myocardial Infarction Genetics (MIGen) Consortium, Geisinger-Regeneron DiscovEHR Collaboration, VA Million Veteran Program, John Concato, J. Michael Gaziano, Christopher J.

O'Donnell, Philip S. Tsao, Sekar Kathiresan, Daniel J. Rader, Peter W. F. Wilson, and Themistocles L. Assimes. Genetics of blood lipids among ˜300,000 multi-ethnic participants of the Million Veteran Program. *Nature Genetics*, 50(11):1514–1523, 2018.

[84] Alberto R Kornblihtt. Chromatin, transcript elongation and alternative splicing. *Nature Structural & Molecular Biology*, 13(1):5–7, 2006.

[85] Katla Kristjánsdóttir, Yeonui Kwak, Nathaniel D. Tippens, John T. Lis, Hyun Min Kang, and Hojoong Kwak. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. Technical report, 2018.

[86] Ivan Krivega and Ann Dean. Enhancer and promoter interactions—long distance calls. *Current Opinion in Genetics & Development*, 22(2):79–85, 2012.

[87] Hojoong Kwak, Nicholas J. Fuda, Leighton J. Core, and John T. Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science (New York, N.Y.)*, 339(6122):950–953, 2013.

[88] Michael L. Lammey, Gary B. Baskin, Andrew P. Gigliotti, D. Rick Lee, John J. Ely, and Meg M. Sleeper. Interstitial myocardial fibrosis in a captive chimpanzee (Pan troglodytes) population. *Comparative Medicine*, 58(4):389–394, 2008.

[89] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

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

[91] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, 2009.

[92] 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.

[93] Y. Li, Y. Sun, Y. Fu, M. Li, G. Huang, C. Zhang, J. Liang, S. Huang, G. Shen, S. Yuan, L. Chen, S. Chen, and A. Xu. Dynamic landscape of tandem 3' UTRs during zebrafish development. *Genome Research*, 22(10):1899–1906, 2012.

[94] 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.

[95] 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.

[96] 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.

[97] Yang Liao, Gordon K Smyth, and Wei Shi. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, 41(10):e108, 2013.

[98] Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–930, 2014.

[99] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–30, 2014.

[100] Lan Lin, Shihao Shen, Peng Jiang, Seiko Sato, Beverly L. Davidson, and Yi Xing. Evolution of alternative splicing in primate brain transcriptomes. *Human Molecular Genetics*, 19(15):2958–2973, 2010.

[101] Yuefeng Lin, Zhihua Li, Fatih Ozsolak, Sang Woo Kim, Gustavo Arango-Argoty, Teresa T. Liu, Scott A. Tenenbaum, Timothy Bailey, A. Paula Monaghan, Patrice M. Milos, and Bino John. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Research*, 40(17):8460–8471, 2012.

[102] Xiaochuan Liu, Jaime Freitas, Dinghai Zheng, Marta S. Oliveira, Mainul Hoque, Torcato Martins, Telmo Henriques, Bin Tian, and Alexandra Moreira. Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in Drosophila melanogaster. *RNA (New York, N.Y.)*, 23(12):1807–1816, 2017.

[103] Liang Ma, Peilin Jia, and Zhongming Zhao. Splicing QTL of human adipose-related traits. *Scientific Reports*, 8(1):318, 2018.

[104] C C MacDonald, J Wilusz, and T Shenk. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Molecular and Cellular Biology*, 14(10):6647–6654, 1994.

[105] Mitchell J. Machiela and Stephen J. Chanock. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)*, 31(21):3555–3557, 2015.

[106] Dig Bijay Mahat, Hojoong Kwak, Gregory T. Booth, Iris H. Jonkers, Charles G. Danko, Ravi K. Patel, Colin T. Waters, Katie Munson, Leighton J. Core, and John T. Lis. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature Protocols*, 11(8):1455–1476, 2016.

[107] Elisa Mariella, Federico Marotta, Elena Grassi, Stefano Gilotto, and Paolo Provero. The Length of the Expressed 3' UTR Is an Intermediate Molecular Phenotype Linking Genetic Variants to Complex Diseases. *Frontiers in Genetics*, 10:714, 2019.

[108] Elisa Mariella, Federico Marotta, Elena Grassi, Stefano Gilotto, and Paolo Provero. The Length of the Expressed 3' UTR Is an Intermediate Molecular Phenotype Linking Genetic Variants to Complex Diseases. *Frontiers in Genetics*, 10, 2019.

[109] Judith Marsman and Julia A. Horsfield. Long distance relationships: Enhancer–promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(11-12):1217–1227, 2012.

[110] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

[111] Andreas Mayer and L. Stirling Churchman. Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing. *Nature Protocols*, 11(4):813–833, 2016.

[112] Andreas Mayer, Julia di Iulio, Seth Maleri, Umut Eser, Jeff Vierstra, Alex Reynolds, Richard Sandstrom, John A. Stamatoyannopoulos, and L. Stirling Churchman. Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell*, 161(3):541–554, 2015.

[113] Andreas Mayer, Heather M Landry, and L Stirling Churchman. Pause & go: From the discovery of RNA polymerase pausing to its functional implications. *Current Opinion in Cell Biology*, 46:72–80, 2017.

[114] Christine Mayr. Evolution and Biological Roles of Alternative 3'UTRs. *Trends in Cell Biology*, 26(3):227–237, 2016.

[115] Christine Mayr. Regulation by 3'-Untranslated Regions. *Annual Review of Genetics*, 51(1):171–194, 2017.

[116] 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.

[117] 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 (New York, N.Y.)*, 342(6159):747–9, 2013.

[118] 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.

[119] modENCODE Consortium, Susan E. Celniker, Laura A. L. Dillon, Mark B. Gerstein, Kristin C. Gunsalus, Steven Henikoff, Gary H. Karpen, Manolis Kellis, Eric C. Lai, Jason D. Lieb, David M. MacAlpine, Gos Micklem, Fabio Piano, Michael Snyder, Lincoln Stein, Kevin P. White, and Robert H. Waterston. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[120] Pamela Moll, Michael Ante, Alexander Seitz, and Torsten Reda. QuantSeq 3' mRNA sequencing for RNA quantification. *Nature Methods*, 11:972, 2014.

[121] Ashleigh E. Moore, Devon M. Chenette, Lauren C. Larkin, and Robert J. Schneider. Physiological networks and disease functions of RNA-binding protein AUF1: RNA-binding protein AUF1. *Wiley Interdisciplinary Reviews: RNA*, 5(4):549–564, 2014.

[122] E. Kathryn Morris, Tancredi Caruso, François Buscot, Markus Fischer, Christine Hancock, Tanja S. Maier, Torsten Meiners, Caroline Müller, Elisabeth Obermaier, Daniel Prati, Stephanie A. Socher, Ilja Sonnemann, Nicole Wäschke, Tesfaye Wubet, Susanne Wurst, and Matthias C. Rillig. Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution*, 4(18):3514–3524, 2014.

[123] Mouse ENCODE Consortium, John A Stamatoyannopoulos, Michael Snyder, Ross Hardison, Bing Ren, Thomas Gingeras, David M Gilbert, Mark Groudine, Michael Bender, Rajinder Kaul, Theresa Canfield, Erica Giste, Audra Johnson, Mia Zhang, Gayathri Balasundaram, Rachel Byron, Vaughan Roach, Peter J Sabo, Richard Sandstrom, A Sandra Stehling, Robert E Thurman, Sherman M Weissman, Philip Cayting, Manoj Hariharan, Jin Lian, Yong Cheng, Stephen G Landt, Zhihai Ma, Barbara J

Wold, Job Dekker, Gregory E Crawford, Cheryl A Keller, Weisheng Wu, Christopher Morrissey, Swathi A Kumar, Tejaswini Mishra, Deepti Jain, Marta Byrska-Bishop, Daniel Blankenberg, Bryan R Lajoie, Gaurav Jain, Amartya Sanyal, Kaun-Bei Chen, Olgert Denas, James Taylor, Gerd A Blobel, Mitchell J Weiss, Max Pimkin, Wulan Deng, Georgi K Marinov, Brian A Williams, Katherine I Fisher-Aylor, Gilberto Desalvo, Anthony Kiralusha, Diane Trout, Henry Amrhein, Ali Mortazavi, Lee Edsall, David McCleary, Samantha Kuan, Yin Shen, Feng Yue, Zhen Ye, Carrie A Davis, Chris Zaleski, Sonali Jha, Chenghai Xue, Alex Dobin, Wei Lin, Meagan Fastuca, Huaien Wang, Roderic Guigo, Sarah Djebali, Julien Lagarde, Tyrone Ryba, Takayo Sasaki, Venkat S Malladi, Melissa S Cline, Vanessa M Kirkup, Katrina Learned, Kate R Rosenbloom, W James Kent, Elise A Feingold, Peter J Good, Michael Pazin, Rebecca F Lowdon, and Leslie B Adams. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, 13(8):418, 2012.

[124] Alisa A Mueller, Tom H Cheung, and Thomas A Rando. All's well that ends well: Alternative polyadenylation and its implications for stem cell biology. *Current Opinion in Cell Biology*, 25(2):222–232, 2013.

[125] Michaela Müller-McNicoll, Oliver Rossbach, Jingyi Hui, and Jan Medenbach. Auto-regulatory feedback by RNA-binding proteins. *Journal of Molecular Cell Biology*, 11(10):930–939, 2019.

[126] Vishal Nanavaty, Elizabeth W. Abrash, Changjin Hong, Sunho Park, Emily E. Fink, Zhuangyue Li, Thomas J. Sweet, Jeffrey M. Bhasin, Srinidhi Singuri, Byron H. Lee, Tae Hyun Hwang, and Angela H. Ting. DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex. *Molecular Cell*, 78(4):752–764.e6, 2020.

[127] Takayuki Nojima, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J. Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J. Proudfoot. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, 161(3):526–540, 2015.

[128] Jung-Min Oh, Chao Di, Christopher C. Venters, Jiannan Guo, Chie Arai, Byung Ran So, Anna Maria Pinto, Zhenxi Zhang, Lili Wan, Ihab Younis, and Gideon Dreyfuss. U1 snRNP telescripting regulates a size–function-stratified human genome. *Nature Structural & Molecular Biology*, 24(11):993–999, 2017.

[129] Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, Robert R. Graham, Arun Manoharan, Ward Ortmann, Tushar Bhangale, Joshua C. Denny, Robert J. Carroll, Anne E. Eyler, Jeffrey D. Greenberg, Joel M. Kremer, Dimitrios A. Pappas, Lei Jiang, Jian Yin, Lingying Ye, Ding-Feng Su, Jian Yang, Gang Xie, Ed Keystone, Harm-Jan Westra, Tõnu Esko, Andres Metspalu, Xuezhong Zhou, Namrata Gupta, Daniel Mirel, Eli A. Stahl, Dorothée Diogo, Jing Cui, Katherine Liao, Michael H.

Guo, Keiko Myouzen, Takahisa Kawaguchi, Marieke J.H. Coenen, Piet L.C.M. van Riel, Mart A.F.J. van de Laar, Henk-Jan Guchelaar, Tom W.J. Huizinga, Philippe Dieudé, Xavier Mariette, S. Louis Bridges, Alexandra Zhernakova, Rene E.M. Toes, Paul P. Tak, Corinne Miceli-Richard, So-Young Bang, Hye-Soon Lee, Javier Martin, Miguel A. Gonzalez-Gay, Luis Rodriguez-Rodriguez, Solbritt Rantapää-Dahlqvist, Lisbeth Ärlestig, Hyon K. Choi, Yoichiro Kamatani, Pilar Galan, Mark Lathrop, Steve Eyre, John Bowes, Anne Barton, Niek de Vries, Larry W. Moreland, Lindsey A. Criswell, Elizabeth W. Karlson, Atsuo Taniguchi, Ryo Yamada, Michiaki Kubo, Jun S. Liu, Sang-Cheol Bae, Jane Worthington, Leonid Padyukov, Lars Klareskog, Peter K. Gregersen, Soumya Raychaudhuri, Barbara E. Stranger, Philip L. De Jager, Lude Franke, Peter M. Visscher, Matthew A. Brown, Hisashi Yamanaka, Tsuneyo Mimori, Atsushi Takahashi, Huji Xu, Timothy W. Behrens, Katherine A. Siminovitch, Shigeki Momohara, Fumihiko Matsuda, Kazuhiko Yamamoto, and Robert M. Plenge. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[130] Katarzyna Oktaba, Wei Zhang, Thea Sabrina Lotz, David Jayhyun Jun, Sandra Beatrice Lemke, Samuel Pak Ng, Emilia Esposito, Michael Levine, and Valérie Hilgers. ELAV Links Paused Pol II to Alternative Polyadenylation in the Drosophila Nervous System. *Molecular Cell*, 57(2):341–348, 2015.

[131] Halit Ongen, Andrew A. Brown, Olivier Delaneau, Nikolaos I. Panousis, Alexandra C. Nica, GTEx Consortium, and Emmanouil T. Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 49(12):1676–1683, 2017.

[132] 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.

[133] Halit Ongen and Emmanouil T. Dermitzakis. Alternative Splicing QTLs in European and African Populations. *American Journal of Human Genetics*, 97(4):567–575, 2015.

[134] Athma A. Pai, Golshid Baharian, Ariane Pagé Sabourin, Jessica F. Brinkworth, Yohann Nédélec, Joseph W. Foley, Jean-Christophe Grenier, Katherine J. Siddle, Anne Dumaine, Vania Yotova, Zachary P. Johnson, Robert E. Lanford, Christopher B. Burge, and Luis B. Barreiro. Widespread Shortening of 3' Untranslated Regions and Increased Exon Inclusion Are Evolutionarily Conserved Features of Innate Immune Responses to Infection. *PLOS Genetics*, 12(9):e1006338, 2016.

[135] Athma A. Pai, Jordana T. Bell, John C. Marioni, Jonathan K. Pritchard, and Yoav Gilad. A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. *PLoS Genetics*, 7(2):e1001316, 2011.

[136] Athma A. Pai, Carolyn E. Cain, Orna Mizrahi-Man, Sherryl De Leon, Noah Lewellen, Jean-Baptiste Veyrieras, Jacob F. Degner, Daniel J. Gaffney, Joseph K. Pickrell,

Matthew Stephens, Jonathan K. Pritchard, and Yoav Gilad. The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels. *PLoS Genetics*, 8(10):e1003000, 2012.

[137] Athma A Pai and Yoav Gilad. Comparative studies of gene regulatory mechanisms. *Current Opinion in Genetics & Development*, 29:68–74, 2014.

[138] Athma A Pai, Jonathan K Pritchard, and Yoav Gilad. The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1):e1004857, 2015.

[139] Robert-Jan Palstra, Bas Tolhuis, Erik Splinter, Rian Nijmeijer, Frank Grosveld, and Wouter de Laat. The β-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics*, 35(2):190–194, 2003.

[140] Zhenhua Pan, Haibo Zhang, Lisa K. Hague, Ju Youn Lee, Carol S. Lutz, and Bin Tian. An intronic polyadenylation site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism. *Gene*, 366(2):325–334, 2006.

[141] Yongjin Park, Abhishek K Sarkar, Liang He, Jose Davila-Velderrain, Philip L De Jager, and Manolis Kellis. A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. Technical report, 2017.

[142] Pedro Patraquim, Maria Warnefors, and Claudio R. Alonso. Evolution of Hox Post-Transcriptional Regulation by Alternative Polyadenylation and MicroRNA Modulation Within 12 Drosophila Genomes. *Molecular Biology and Evolution*, 28(9):2453–2460, 2011.

[143] Bryan J. Pavlovic, Lauren E. Blake, Julien Roux, Claudia Chavarria, and Yoav Gilad. A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Scientific Reports*, 8(1):15312, 2018.

[144] 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.

[145] Brandon L. Pierce, Lin Tong, Maria Argos, Kathryn Demanelis, Farzana Jasmine, Muhammad Rakibuz-Zaman, Golam Sarwar, Md. Tariqul Islam, Hasan Shahriar, Tariqul Islam, Mahfuzar Rahman, Md. Yunus, Muhammad G. Kibriya, Lin S. Chen, and Habibul Ahsan. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nature Communications*, 9(1):804, 2018.

[146] Pedro A B Pinto, Telmo Henriques, Marta O Freitas, Torcato Martins, Rita G Domingues, Paulina S Wyrzykowska, Paula A Coelho, Alexandre M Carmo, Claudio E Sunkel, Nicholas J Proudfoot, and Alexandra Moreira. RNA polymerase II kinetics in *polo* polyadenylation signal selection: RNA polymerase II kinetics and alternative polyadenylation. *The EMBO Journal*, 30(12):2431–2444, 2011.

[147] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[148] David H. Price. Transient pausing by RNA polymerase II. *Proceedings of the National Academy of Sciences*, 115(19):4810–4812, 2018.

[149] Nick J. Proudfoot. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (New York, N.Y.)*, 352(6291):aad9926, 2016.

[150] Nick J. Proudfoot, Andre Furger, and Michael J. Dye. Integrating mRNA Processing with Transcription. *Cell*, 108(4):501–512, 2002.

[151] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–504, 2005.

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

[153] Peter B. Rahl, Charles Y. Lin, Amy C. Seila, Ryan A. Flynn, Scott McCuine, Christopher B. Burge, Phillip A. Sharp, and Richard A. Young. C-Myc regulates transcriptional pause release. *Cell*, 141(3):432–445, 2010.

[154] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, 2016.

[155] Tommer Ravid and Mark Hochstrasser. Diversity of degradation signals in the ubiquitin–proteasome system. *Nature Reviews Molecular Cell Biology*, 9(9):679–689, 2008.

[156] Kirsten A. Reimer, Claudia Mimoso, Karen Adelman, and Karla M. Neugebauer. Rapid and Efficient Co-Transcriptional Splicing Enhances Mammalian Gene Expression. Technical report, 2020.

[157] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[158] David W. Rogers, Marvin A. Böttcher, Arne Traulsen, and Duncan Greig. Ribosome reinitiation can explain length-dependent translation of messenger RNA. *PLOS Computational Biology*, 13(6):e1005592, 2017.

[159] Irene Gallego Romero, Shyam Gopalakrishnan, and Yoav Gilad. Widespread conservation of chromatin accessibility patterns and transcription factor binding in human and chimpanzee induced pluripotent stem cells. Technical report, 2018.

[160] Jason Sadek, Amr Omer, Derek Hall, Kholoud Ashour, and Imed Eddine Gallouzi. Alternative polyadenylation and the stress response. *Wiley Interdisciplinary Reviews: RNA*, page e1540, 2019.

[161] R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, and C. B. Burge. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, 320(5883):1643–1647, 2008.

[162] Benjamin J. Schmiedel, Divya Singh, Ariel Madrigal, Alan G. Valdovino-Gonzalez, Brandie M. White, Jose Zapardiel-Gonzalo, Brendan Ha, Gokmen Altay, Jason A. Greenbaum, Graham McVicker, Grégory Seumois, Anjana Rao, Mitchell Kronenberg, Bjoern Peters, and Pandurangan Vijayanand. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*, 175(6):1701–1715.e16, 2018.

[163] Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan M. Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, Evan E. Eichler, and Deanna M. Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017.

[164] Sarah Sheppard, Nathan D. Lawson, and Lihua Julie Zhu. Accurate identification of polyadenylation sites from 3' end deep sequencing using a naïve Bayes classifier. *Bioinformatics*, 29(20):2564–2571, 2013.

[165] Yongsheng Shi. Alternative polyadenylation: New insights from global analyses. *RNA (New York, N.Y.)*, 18(12):2105–2117, 2012.

[166] Yoichiro Shibata, Nathan C. Sheffield, Olivier Fedrigo, Courtney C. Babbitt, Matthew Wortham, Alok K. Tewari, Darin London, Lingyun Song, Bum-Kyu Lee, Vishwanath R. Iyer, Stephen C. J. Parker, Elliott H. Margulies, Gregory A. Wray, Terrence S. Furey, and Gregory E. Crawford. Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. *PLoS Genetics*, 8(6):e1002789, 2012.

[167] David A. Siegel, Olivier Le Tonqueze, Anne Biton, Noah Zaitlen, and David J. Erle. Massively Parallel Analysis of Human 3′ UTRs Reveals that AU-Rich Element Length and Registration Predict mRNA Destabilization. Technical report, 2020.

[168] 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):1716, 2018.

[169] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, 2017.

[170] Lingyun Song and Gregory E. Crawford. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb.prot5384, 2010.

[171] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, and Emmanouil T Dermitzakis. Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224, 2007.

[172] B. J. Strober, R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019.

[173] Bhairavi Swaminathan, Gumar Thorleifsson, Magnus Jöud, Mina Ali, Ellinor Johnsson, Ram Ajore, Patrick Sulem, Britt-Marie Halvarsson, Gumundur Eyjolfsson, Vilhelmina Haraldsdottir, Christina Hultman, Erik Ingelsson, Sigurur Y. Kristinsson, Anna K. Kähler, Stig Lenhoff, Gisli Masson, Ulf-Henrik Mellqvist, Robert Månsson, Sven Nelander, Isleifur Olafsson, Olof Sigurardottir, Hlif Steingrimsdóttir, Annette Vangsted, Ulla Vogel, Anders Waage, Hareth Nahi, Daniel F. Gudbjartsson, Thorunn Rafnar, Ingemar Turesson, Urban Gullberg, Kári Stefánsson, Markus Hansson, Unnur Thorsteinsdóttir, and Björn Nilsson. Variants in *ELL2* influencing immunoglobulin levels associate with multiple myeloma. *Nature Communications*, 6:7213, 2015.

[174] Y Takagaki, L C Ryner, and J L Manley. Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes & Development*, 3(11):1711–1724, 1989.

[175] 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*, 8:14519, 2017.

[176] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[177] The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019.

[178] 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.

[179] Bin Tian and James L. Manley. Alternative polyadenylation of mRNA precursors. *Nature Reviews. Molecular Cell Biology*, 18(1):18–30, 2017.

[180] Bin Tian, Zhenhua Pan, and Ju Youn Lee. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Research*, 17(2):156–165, 2007.

[181] Jianbo Tian, Zhihua Wang, Shufang Mei, Nan Yang, Yang Yang, Juntao Ke, Ying Zhu, Yajie Gong, Danyi Zou, Xiating Peng, Xiaoyang Wang, Hao Wan, Rong Zhong, Jiang Chang, Jing Gong, Leng Han, and Xiaoping Miao. CancerSplicingQTL: A database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Research*, 47(D1):D909–D916, 2019.

[182] Bas Tolhuis, Robert-Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell*, 10(6):1453–1465, 2002.

[183] Jacob M. Tome, Nathaniel D. Tippens, and John T. Lis. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nature Genetics*, 50(11):1533–1541, 2018.

[184] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2):124–30, 2013.

[185] 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.

[186] Nissi Varki, Dan Anderson, James G. Herndon, Tho Pham, Christopher J. Gregg, Monica Cheriyan, James Murphy, Elizabeth Strobert, Jo Fritz, James G. Else, and Ajit Varki. ORIGINAL ARTICLE: Heart disease is common in humans and chimpanzees, but is caused by different pathological processes: Heart disease in hominids. *Evolutionary Applications*, 2(1):101–112, 2009.

[187] Shobha Vasudevan, Stuart W. Peltz, and Carol J. Wilusz. Non-stop decay—a new mRNA surveillance pathway. *BioEssays*, 24(9):785–788, 2002.

[188] Chris Wallace, Maxime Rotival, Jason D. Cooper, Catherine M. Rice, Jennie H. M. Yang, Mhairi McNeill, Deborah J. Smyth, David Niblett, François Cambien, Laurence Tiret, John A. Todd, David G. Clayton, and Stefan Blankenberg. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics*, 21(12):2815–2824, 2012.

[189] Ruijia Wang, Ram Nambiar, Dinghai Zheng, and Bin Tian. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Research*, 46(D1):D315–D319, 2018.

[190] Ruijia Wang, Dinghai Zheng, Ghassan Yehia, and Bin Tian. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research*, page gr.237826.118, 2018.

[191] Sidney H. Wang, Chiaowen Joyce Hsiao, Zia Khan, and Jonathan K. Pritchard. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biology*, 19(1):83, 2018.

[192] Michelle C. Ward and Yoav Gilad. Cracking the regulatory code. *Nature*, 550(7675):190–191, 2017.

[193] Michelle C Ward and Yoav Gilad. A generally conserved response to hypoxia in iPSC-derived cardiomyocytes from humans and chimpanzees. *eLife*, 8:e42374, 2019.

[194] Michael Weber, Ines Hellmann, Michael B Stadler, Liliana Ramos, Svante Pääbo, Michael Rebhan, and Dirk Schübeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39(4):457–466, 2007.

[195] Erin M. Wissink, Anniina Vihervaara, Nathaniel D. Tippens, and John T. Lis. Nascent RNA analyses: Tracking transcription and its regulation. *Nature Reviews Genetics*, 20(12):705–723, 2019.

[196] Gregory A. Wray. The evolutionary significance of cis-regulatory mutations. *Nature Reviews. Genetics*, 8(3):206–216, 2007.

[197] Zhengrong Xing, Peter Carbonetto, and matthew Stephens. Flexible signal denoising via flexible empirical Bayes shrinkage. *arXiv*, 1605.07787, 2016.

[198] A. Yamashita and O. Takeuchi. Translational control of mRNAs by 3'-Untranslated region binding proteins. *BMB reports*, 50(4):194–200, 2017.

[199] Yanbo Yang, Qiong Zhang, Ya-Ru Miao, Jiajun Yang, Wenqian Yang, Fangda Yu, Dongyang Wang, An-Yuan Guo, and Jing Gong. SNP2APA: A database for evaluating effects of genetic variants on alternative polyadenylation in human cancers. *Nucleic Acids Research*, 2019.

[200] Yanbo Yang, Qiong Zhang, Ya-Ru Miao, Jiajun Yang, Wenqian Yang, Fangda Yu, Dongyang Wang, An-Yuan Guo, and Jing Gong. SNP2APA: A database for evaluating effects of genetic variants on alternative polyadenylation in human cancers. *Nucleic Acids Research*, 48(D1):D226–D232, 2020.

[201] Chen Yao, George Chen, Ci Song, Joshua Keefe, Michael Mendelson, Tianxiao Huan, Benjamin B. Sun, Annika Laser, Joseph C. Maranville, Hongsheng Wu, Jennifer E. Ho, Paul Courchesne, Asya Lyass, Martin G. Larson, Christian Gieger, Johannes Graumann, Andrew D. Johnson, John Danesh, Heiko Runz, Shih-Jen Hwang, Chunyu Liu, Adam S. Butterworth, Karsten Suhre, and Daniel Levy. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, 9(1):1–11, 2018.

[202] Peng Yao, Alka A. Potdar, Abul Arif, Partho Sarothi Ray, Rupak Mukhopadhyay, Belinda Willard, Yichi Xu, Jun Yan, Gerald M. Saidel, and Paul L. Fox. Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell*, 149(1):88–100, 2012.

[203] Gene Yeo and Christopher B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 11(2-3):377–394, 2004.

[204] Oh Kyu Yoon, Tiffany Y. Hsu, Joo Hyun Im, and Rachel B. Brem. Genetics and Regulatory Impact of Alternative Polyadenylation in Human B-Lymphoblastoid Cells. *PLOS Genetics*, 8(8):e1002882, 2012.

[205] Julia Zeitlinger, Alexander Stark, Manolis Kellis, Joung-Woo Hong, Sergei Nechaev, Karen Adelman, Michael Levine, and Richard A. Young. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature Genetics*, 39(12):1512–1516, 2007.

[206] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, 2014.

[207] Xiang Zhou, Carolyn E. Cain, Marsha Myrthil, Noah Lewellen, Katelyn Michelini, Emily R. Davenport, Matthew Stephens, Jonathan K. Pritchard, and Yoav Gilad. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biology*, 15(12):547, 2014.