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

Copyright © 2020 by Briana Erin Mittleman

All Rights Reserved

Freely available under a CC-BY 4.0 International license

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

ABSTRACT

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

CHAPTER 1

INTRODUCTION

1.1 Human Genetics and the search for the genetic basis of human phenotypes

1.2 Functional genomics and the investigation of the non-coding regions of the genome

1.3 Tuberculosis and the genetic basis of susceptibility

1.4 Single cell sequencing technology and the future of functional genomics

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 [11] 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 [3, 10]. 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) [3]. 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 [44]. 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 [81, 56, 76]. 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 [57, 81]), 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 [40, 77] through the translation of truncated mRNAs into truncated proteins, which impairs tumour-suppressive functions [40, 77].

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 [42, 93, 91, 6, 52]. While using existing RNA-seq to study APA is economical, identifying PAS and estimating usage using RNA-seq are error-prone and often imprecise [25]. 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 [15, 58, 44, 69]. 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 [15, 58, 44, 69]. 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 [75, 80]. 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 [81, 56, 76]. 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) [77]. 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 [33, 5, 63]. 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 [92] 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) [82]. 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 [33, 5, 63].

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 (Supplementart 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) [87]. 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 [43] and tested *cis*-associations between genetic variants and PAS usage, correcting for batch and the top principal components (Methods, Supplementary Figure 2.11,Supplementary Figure 2.12,Supplementary Figure 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 (Fig-

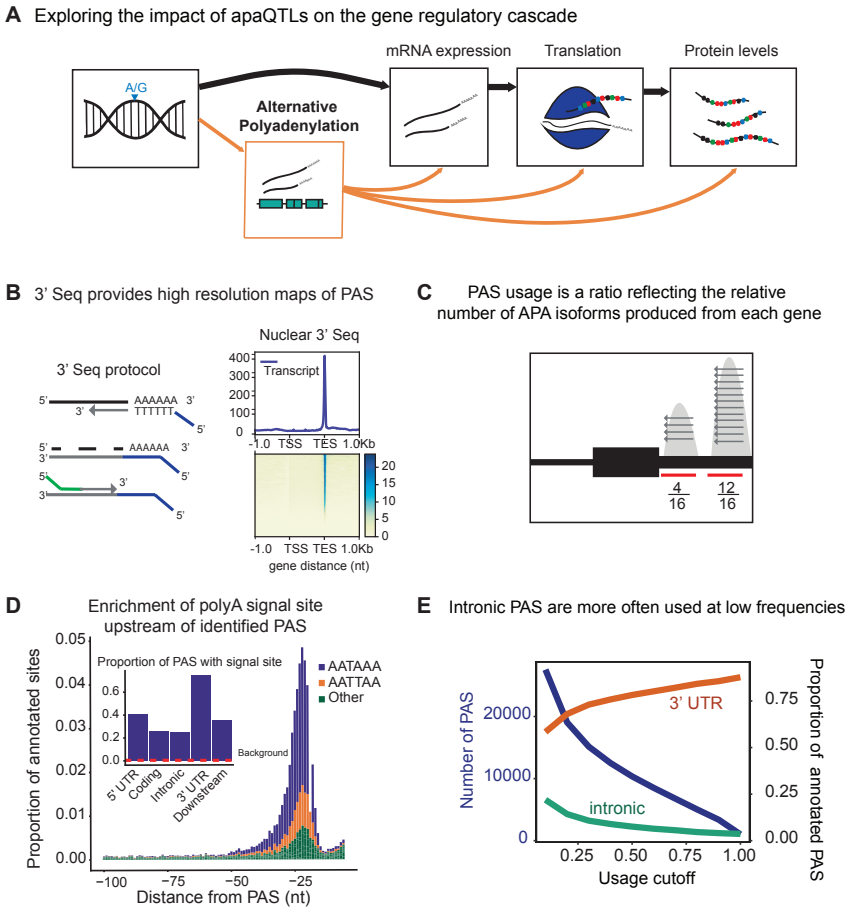

Figure 2.1: **(A)** Schematic of how genetic variants affect phenotypes by percolating through gene regulatory layers (black arrows). We aimed to understand how genetic variation can mediate gene regulation through alternative polyadenylation (orange arrows). **(B)** *(Left)* Schematic of Lexogen Reverse Quant Seq protocol for 3' Sequencing [61] *(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.

ure 2.2B, Supplementary Figure 2.14). The 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 2.14).

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

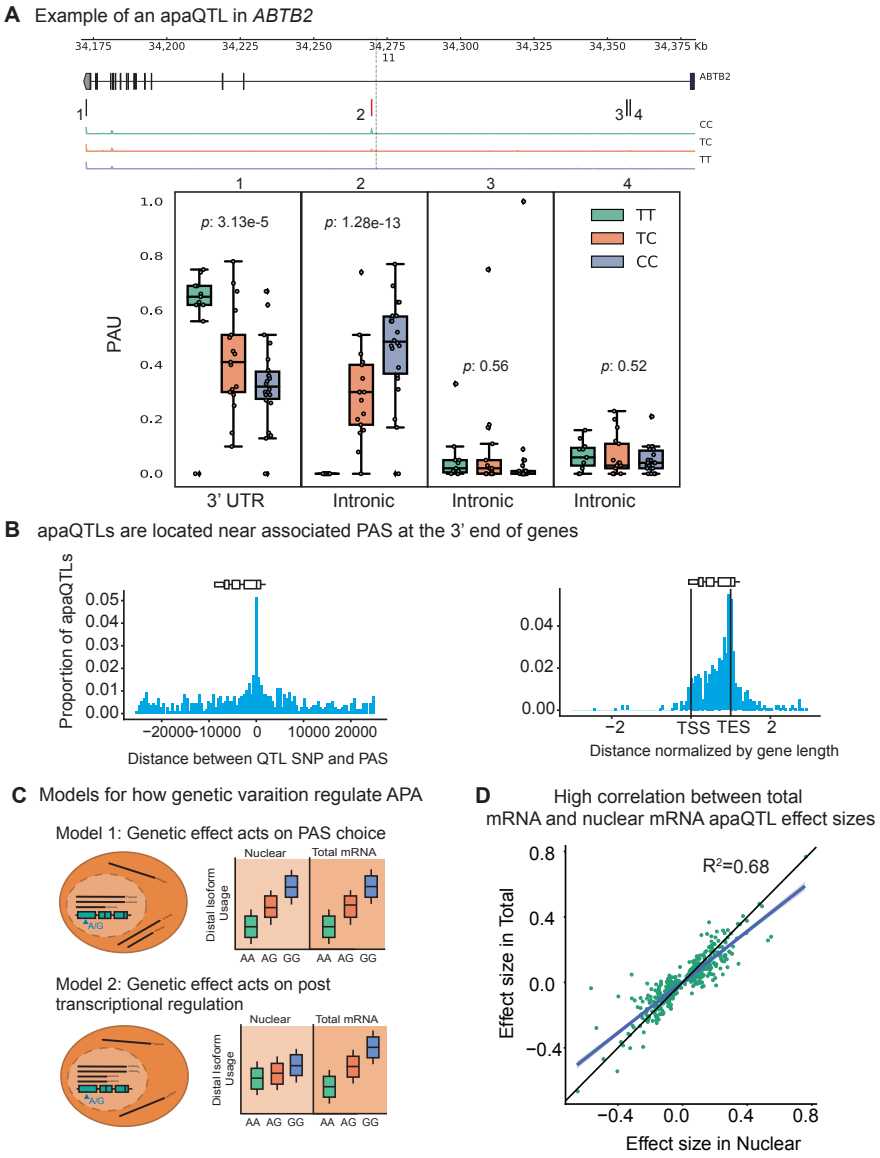

Figure 2.2: **(A)** An apaQTL in the *ABTB2* gene impact usage of an intronic PAS. *(Top)* Gene track and identified PAS. Each bar represents a potential isoform. The red bar corresponds to the isoform most strongly associated with the apaQTL. The vertical dotted line represents the position of the lead 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.

isoforms with premature PAS that are subject to rapid decay. Of interest, we found that 13 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 [44]. 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 transcrip-

tion 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 [44, 3]. 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 [18]. 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; $p = 0.0006$) compared to genes with no molecular association (Figure 2.4A, Supplementary Figure 2.22) [44, 3]. 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.

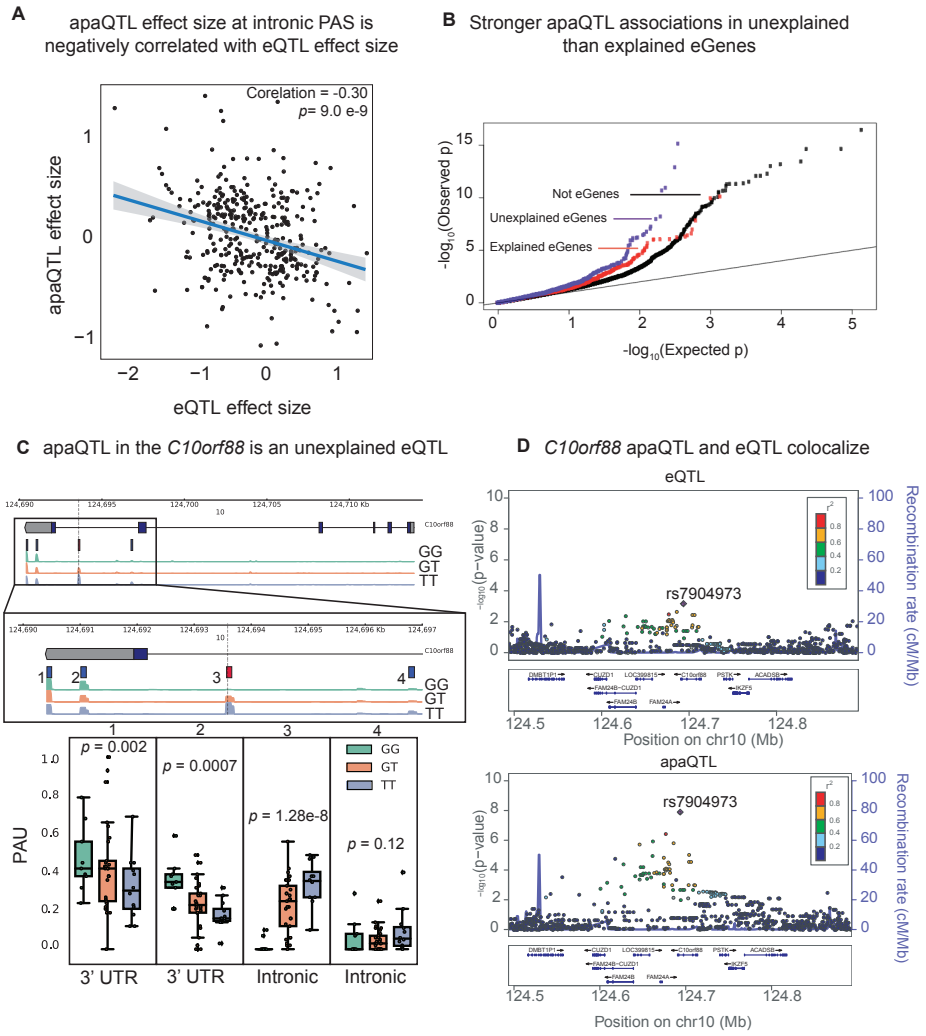

Figure 2.3: **(A)** Scatter plot of intronic apaQTL effect sizes plotted against their eQTL effect sizes shows negative correlation. **(B)** Quantile-quantile (Q-Q) plot for apaQTLs shows that apaQTLs are more highly enriched in unexplained eGenes (purple dots) compared to explained eGenes (red dots). **(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 [36].

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 [3]. 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 [57], 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 [90], or differential rates of translation re-initiation at the end of a translation cycle [74].

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 [3, 10]. 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) [34]. For example, an apaQTL that colocalizes with the eQTLs in the C10orf88 gene (rs7904973) has been associated with increased LDL cholesterol [36], 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 [30, 40]. 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 [79]. Interestingly, multiple myeloma is among the cancer types in which widespread dysregulation of intronic APA has been documented previously [77, 40].

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

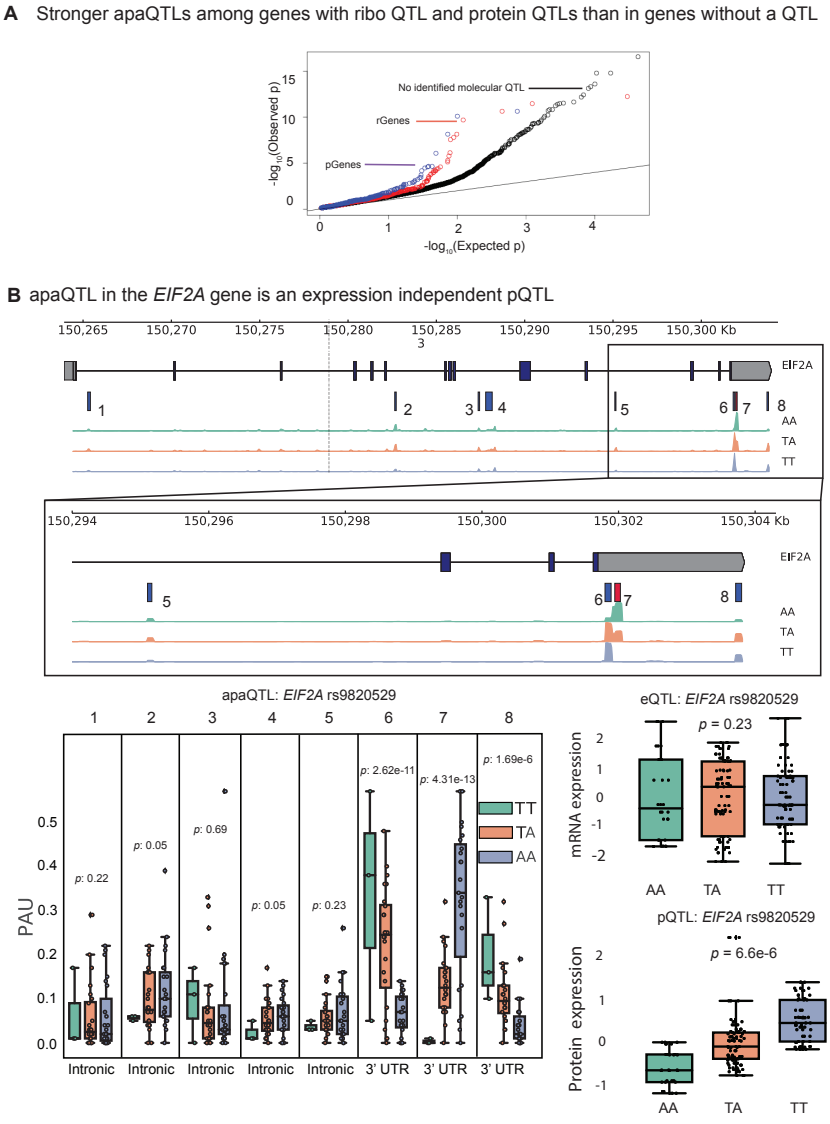

Figure 2.4: **(A)** Quantile-quantile (Q-Q) plot for apaQTLs separated by genes in previously detected rQTLS (red) and pQTLs (purple) that are not eQTLs. Black points are apaQTL genes with no pQTL, rQTL, or eQTL. **(B)***(Top)* Gene track and identified PAS in the *EIF2A* gene. The red bar corresponds 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 [42, 93, 91], 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) [87].

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) [42]. 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 [63], and RNAP II pausing [20]. Indeed, recent studies have reported that sequence and chromatin context can pause or slow down RNAP II elongation across the gene body [54], suggesting that variation in RNAPII pausing may impact PAS choice [20]. For example, in Drosophila melanogaster, paused RNAPII promotes the recruitment of ELAV on the pre-mRNA, which prevents usage of a proximal PAS [65]. 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 [49]. 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 [81, 85]. 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 [27].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 [53]. 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 [53]. 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 [61] 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 [12] using STAR RNA-seq aligner [17] using default settings with the WASP mode to filter out reads mapping with allelic bias [84]. 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 [80, 75, 4]. We verified the individual identity of all bam files using VerifyBamID [32]. 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,Supplementary Figure 2.27,Supplementary Figure 2.28,Supplementary Figure 2.29).

2.5.5 Identification and characerization 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 [34]. When a region mapped to multiple genes we used a hierarchical model, similar to the method used by Lin et al. [48] 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 [87](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) [39]. We calculated the 5' splice site strength using the MaxEntScore tool, for each of the introns in our annotation [92]. 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 [26]. 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 [4]. 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 [4].

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 [47]. We ran the leafcutter_ds.R script on chromosomes 1-22 separately using the cellular fraction label as the sample group identifier [43]. 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[44], available at http://eqtl.uchicago.edu/jointLCL/genotypesYRI.gen.txt.gz [44]. 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 [67]. 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 [34]. We overlapped the eQTLs identified and published in Li et al. 2016[44] 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 [44]. 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[44] 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 [86]. 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) [66].

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[3]. 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 [44]. We used normalized RNA expression, RiboSeq values, and protein levels, published in Li et al. 2016 [44].

2.5.13 PAS heritability estimates and apaQTL overlap with GWAS Catalog

PAS heritability estimates and apaQTL overlap with GWAS Catalog

We downloaded GWAS summary statistics from both Astle et al. and Okada et al. [2, 64] We augmented our PAS sites by 500bp on either side and ran LD score regression using methods described in Bulik-Sullivan et al. [8] We downloaded the CRCh37hg19 GWAS catalog for UCSC table browser [34]. We identified SNPs in LD with the nuclear apaQTLs using the LDproxy tool from LDlink with YRI as the population [50]. 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

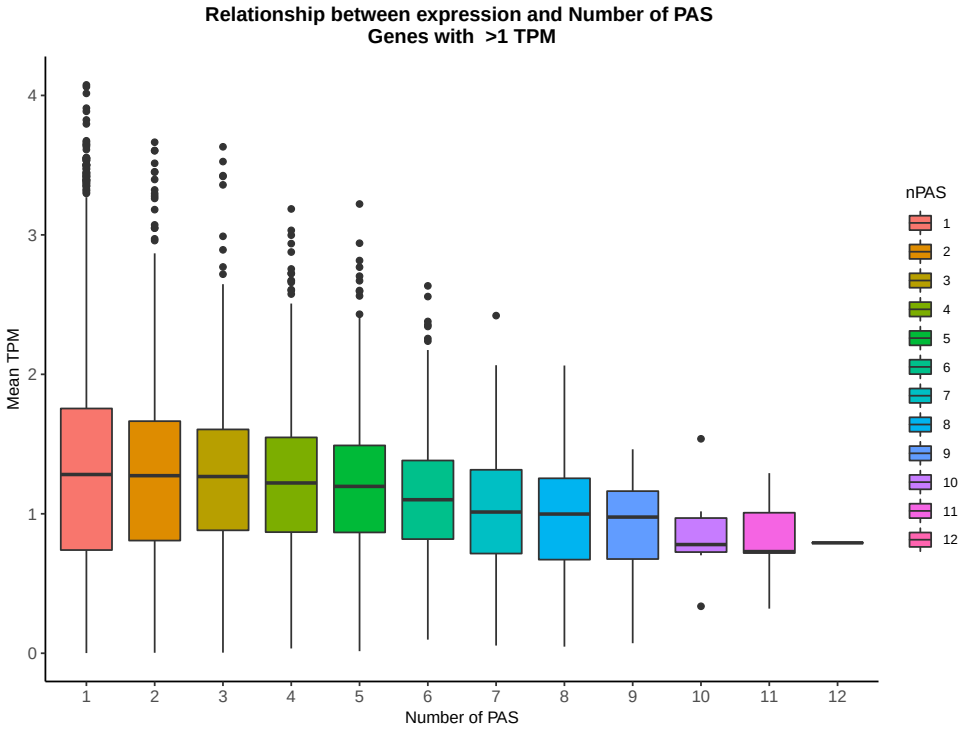

Figure 2.5: **Relationship between Number of PAS and gene expression** Relationship between number of PAS identified in our study and gene expression levels (TPM) as measured from GEUVADIS YRI LCLs [39]. Genes with mean TPM < 1 across individuals were considered not expressed and thus were removed for this analysis.

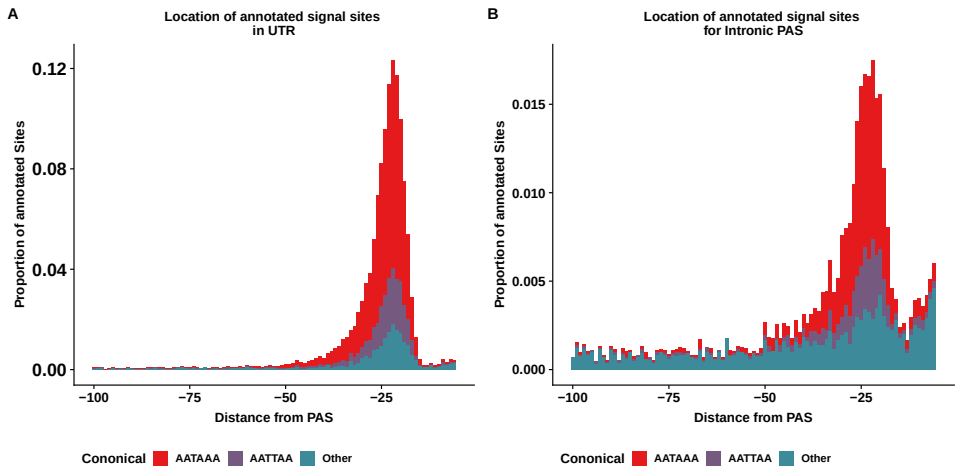

Figure 2.6: **Distribution of signal sites upstream of PAS. Supplement to Figure 2.1D (A)** Stacked density plot showing the signal site distribution for PAS in 3' UTR. Other signal sites are AAAAAA, AAAAAG, AATACA, AATAGA, AATATA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAAA. **(B)** Stacked density plot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

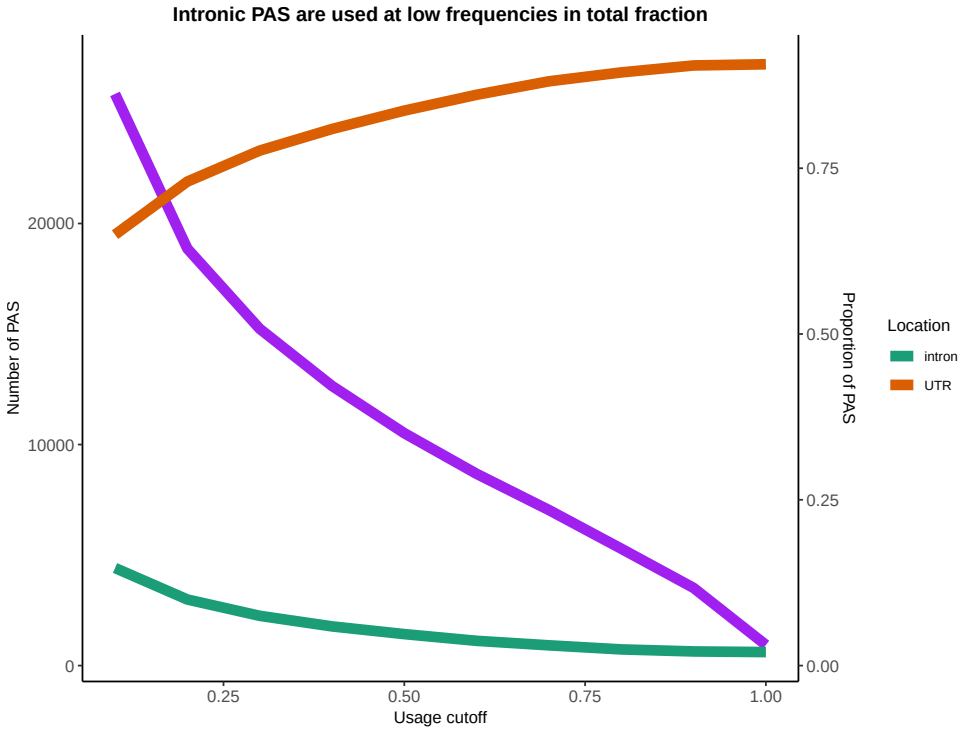

Figure 2.7: **Proportion of PAS in 3' UTRs and introns as predicted from total 3' Seq. Additional figures corresponding to Figure 2.1E.** Number of PAS identified with usage larger than the usage cutoff (x axis) in the total mRNA fraction (purple). Proportion of PAS in introns when PAS are filtered by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

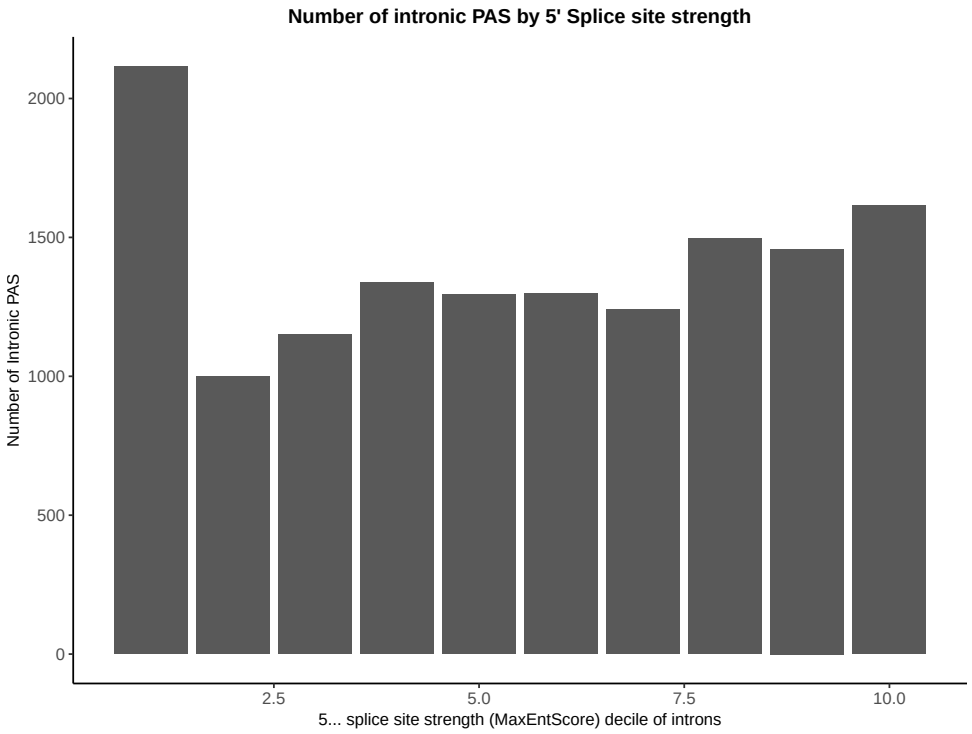

Figure 2.8: **Intronic PAS 5' Splice site strength** Intronic PAS are enriched in introns with the weakest 5' splice sites. Splice site strengths for all introns were calculated using MaxEntScore [92]

.

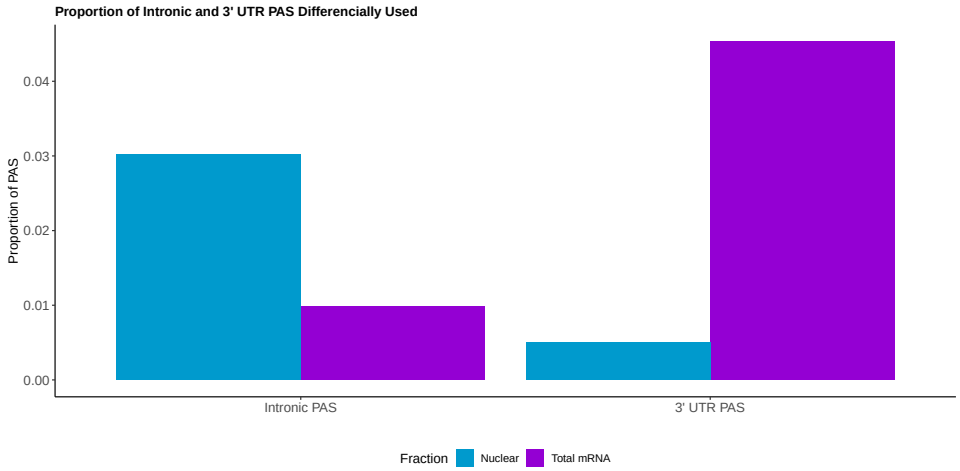

Figure 2.9: **Location of PAS differentially used** We adapted LeafCutter to identify genes with significant differential usage of PAS between the total and nuclear fraction. The majority of PAS preferentially used in the nuclear fraction are intronic, whereas the majority of PAS preferentially used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

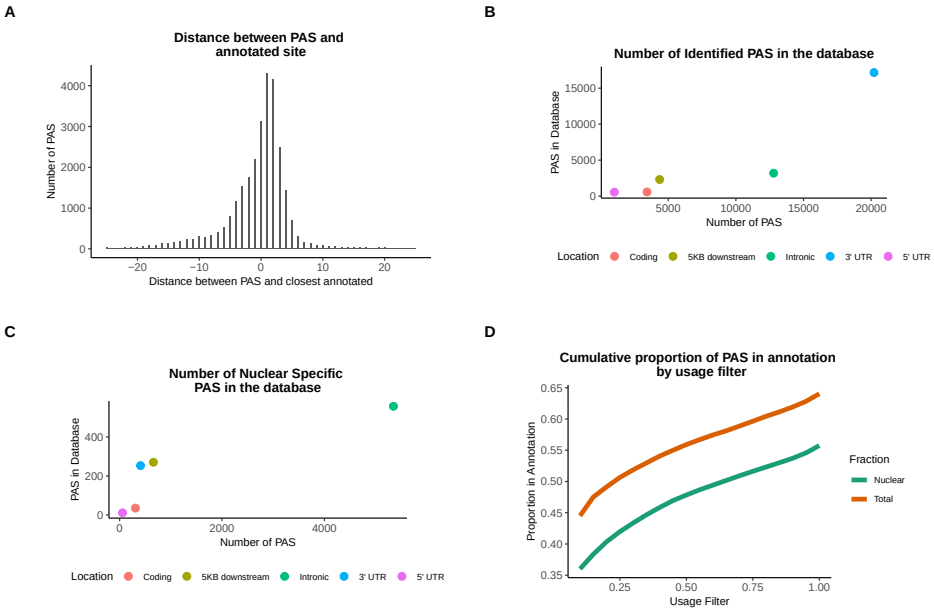

Figure 2.10: **Comparison of our 3' Seq PAS to previous PAS annotations (A)** Distribution of distance between PAS and closest annotated site in the annotation database (PolyA_DB release 3.2) [87]. **(B)** Scatter plot showing the number of PAS we identified in our study (X axis) versus the number of 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.

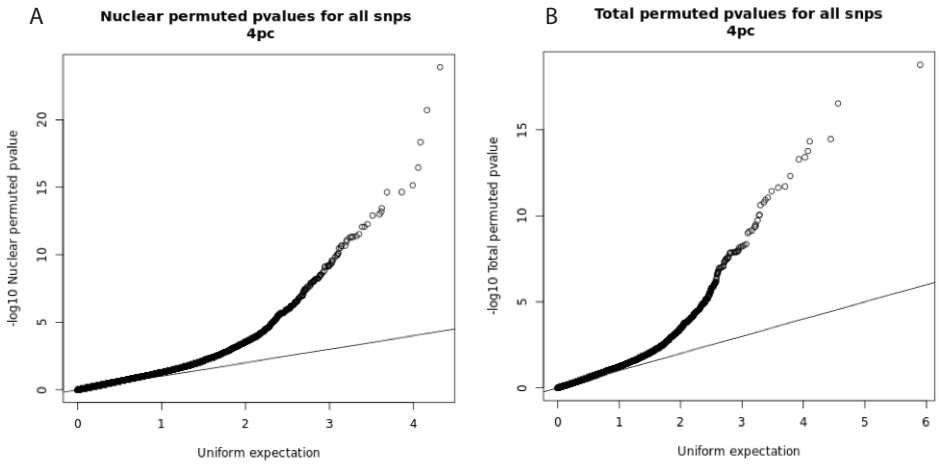

Figure 2.11: **Q-Q plots for apaQTLs (A)** Q-Q plot for nuclear apaQTLs, plotting adjusted p-values of the top SNP PAS associations. **(B)** Q-Q plot for total apaQTLs, plotting adjusted p-values of the top SNP PAS associations.

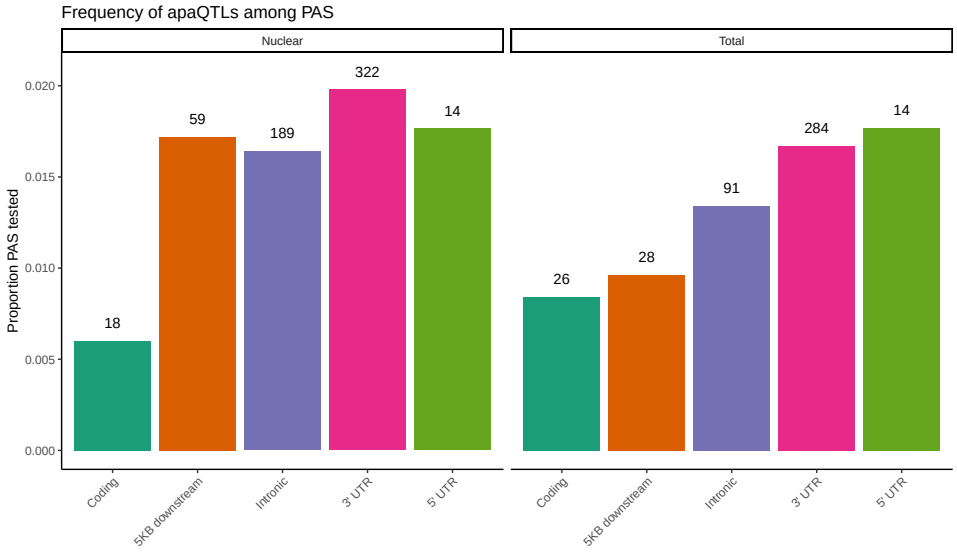

Figure 2.12: **Proportion of PAS tested with an apaQTL** Proportion of PAS in different genomic locations with a significant apaQTL. The numbers above each bar represent the number of identified apaQTL for each location.

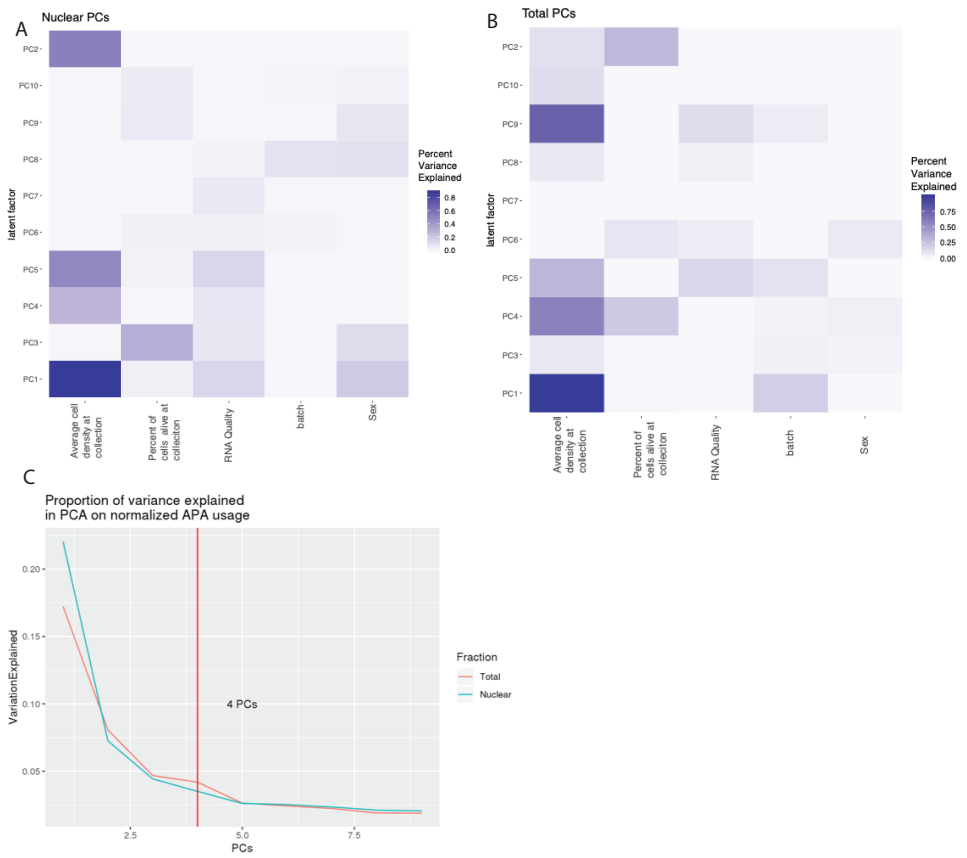

Figure 2.13: **Analysis of the PCs of APA usage (A)** Proportion of variance explained in the 10 first PCs by experimental variables in nuclear APA usage. We used a linear model to look at correlation between PC and each covariate. **(B)** Proportion of variance explained in the 10 first PCs by experimental 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.

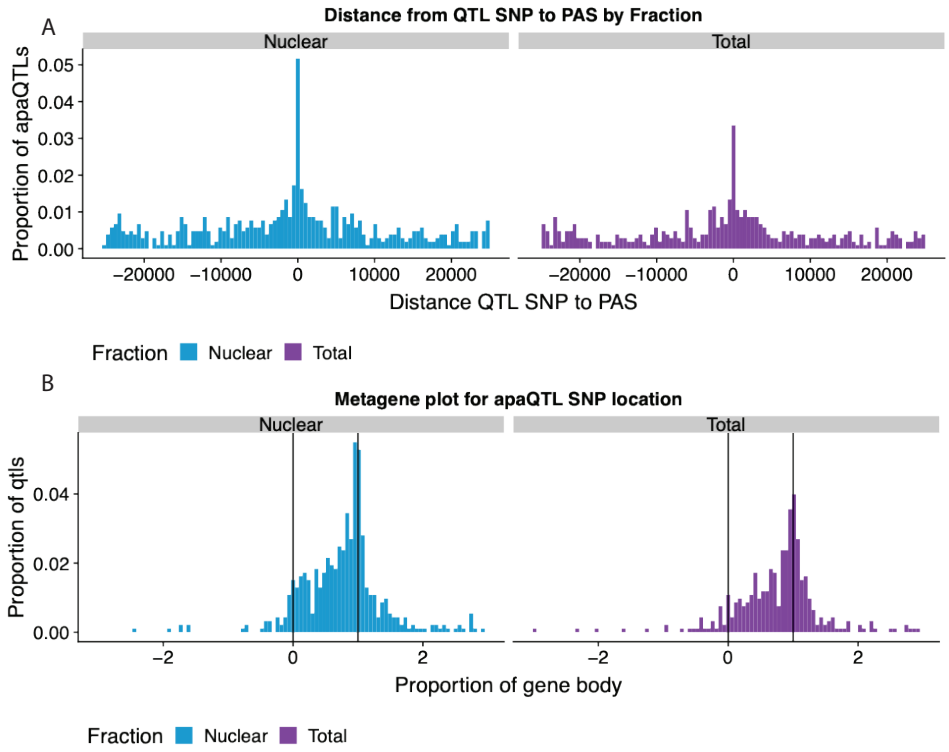

Figure 2.14: **apaQTLs in both fractions are associated with PAS near SNP and at the transcription end site. Supplement to Figure 2.2B and 2.2C. (A)** Histogram showing the distribution of the distance between lead apaQTL SNP and the PAS, separated by mRNA fraction. **(B)** Histogram showing the distribution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

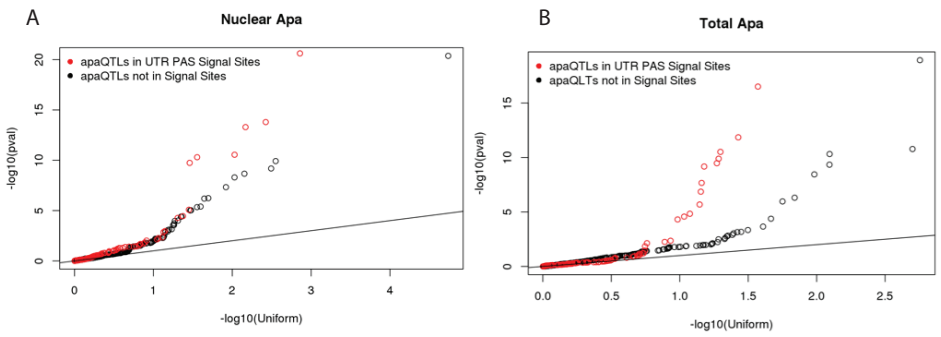

Figure 2.15: **Signal site disruption (A)** Q-Q plot showing the nuclear apaQTL p-values for SNP in signal sites upstream of 3' UTR PAS compared to matched SNPs (equal distance) upstream of a set of 3' UTR PAS without identified signal sites. **(B)** Similar to panel A, but for total apaQTLs.

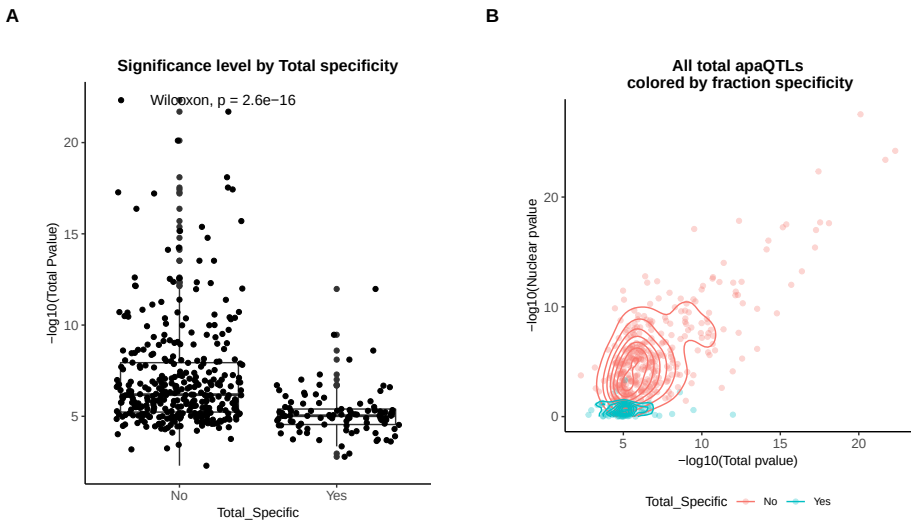

Figure 2.16: **Total mRNA specific apaQTLs show weaker association than do shared apaQTLs (A)** Boxplot showing the -log10(p-value) of the nominal total apaQTL associations separated by whether the association is also identified in the nuclear mRNA fraction. ApaQTLs that are total-specific have significantly 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.

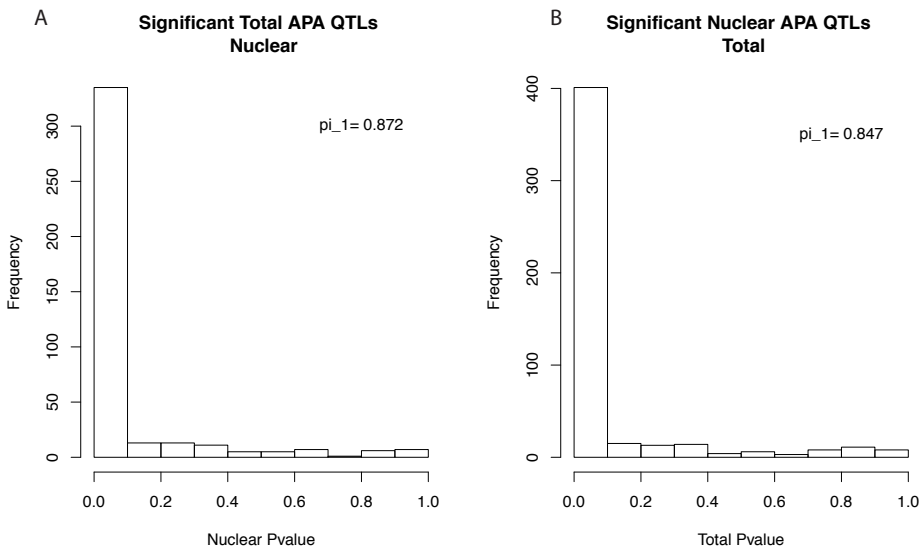

Figure 2.17: **apaQTL sharing between fractions (A)** Histogram showing the P-value distribution of the apaQTL associations between the lead total apaQTL SNP and the corresponding PAS ascertained using our 3'-Seq data from the nuclear mRNA fraction. Values were calculated based on PAS tested in both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

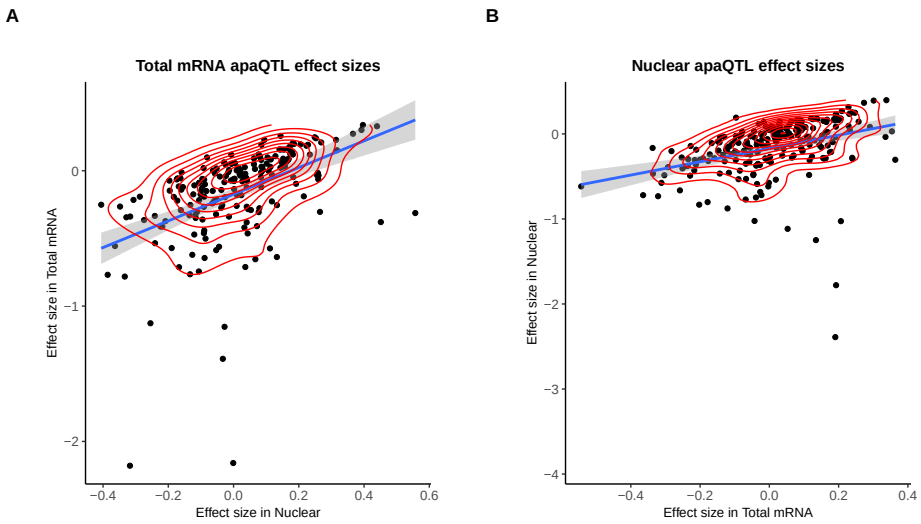

Figure 2.18: **Correlation of effect sizes for apaQTLs discovered in total and nuclear mRNA fractions (A)** Normalized effect sizes ascertained in total mRNA and nuclear fraction of total apaQTLs tested in both fractions. **(B)** Normalized effect sizes ascertained in total mRNA and nuclear fraction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

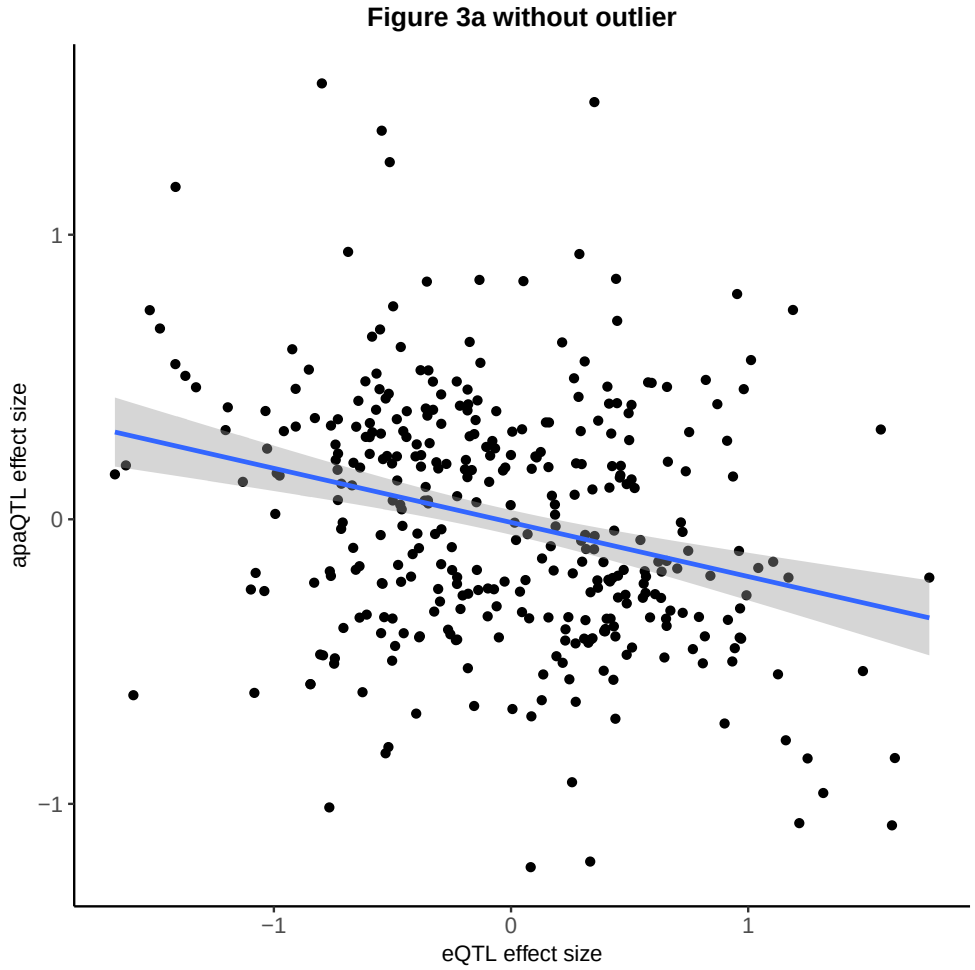

Figure 2.19: **Figure 2.3A without outlier SNP** Scatter plot showing the relationship between intronic nuclear apaQTL effect size and eQTL effect size after removing outlier SNPs (Filtered for SNPs with eQTL effect size < -2.0).

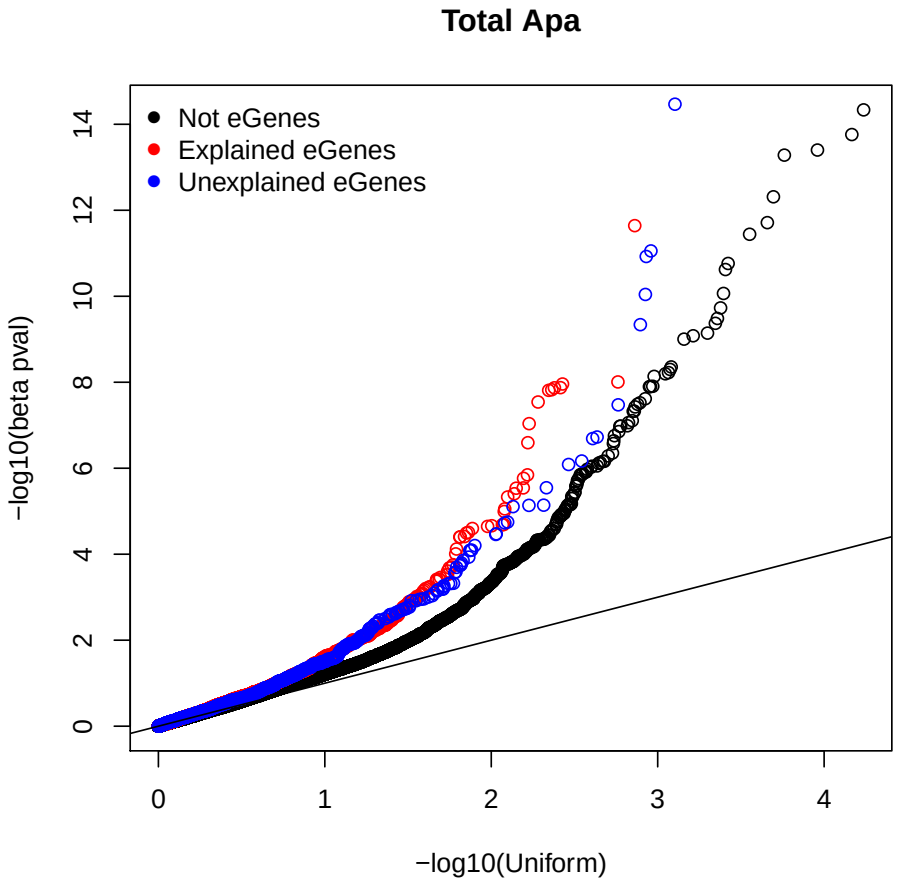

Figure 2.20: **Overlap between apaQTLs in total fraction and eQTLs, supplement to Figure 2.3B** QQ-plot showing the total apaQTL (adjusted) p-values separated by whether the gene harbors an explained (red) or unexplained (blue) eQTLs. We observe an enrichment for low apaQTL association p-values in genes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

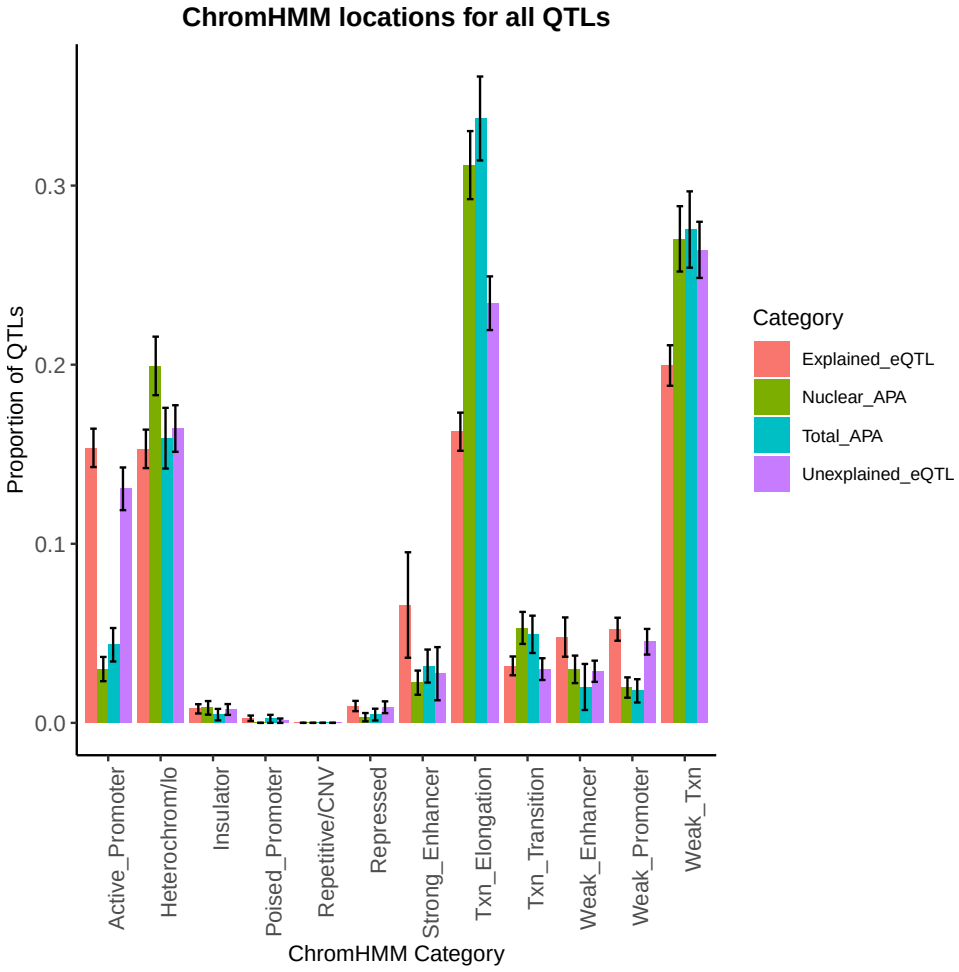

Figure 2.21: **Proportion of apaQTLs and eQTLs by Chromatin state** Bar plot showing the proportion of apaQTLs located in each of the 12 chromatin states from chromHMM. We find that the location profile of apaQTLs is more similar to that of unexplained eQTLs than that of explained eQTLs. Error bars represent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

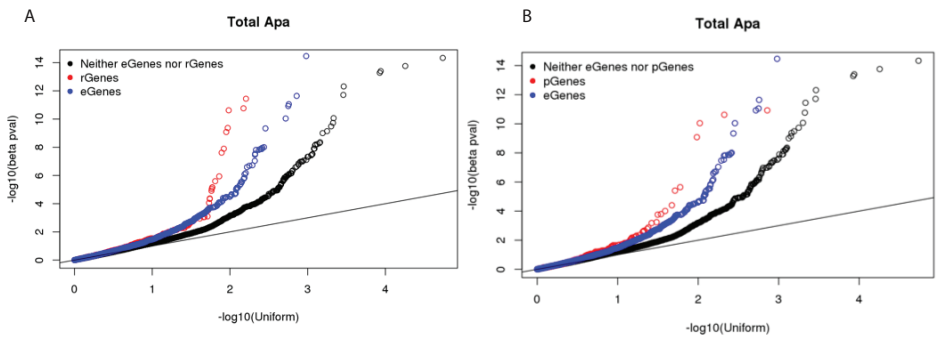

Figure 2.22: **Overlap between apaQTLs in total fraction and eQTLs, rQTLs and pQTLs supplement to Figure 2.4A (A)** QQ-plot showing the total apaQTL (adjusted) p-values separated by whether the corresponding gene has a ribosome occupancy QTL (red) or an eQTL (red). We see an enrichment for low apaQTL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

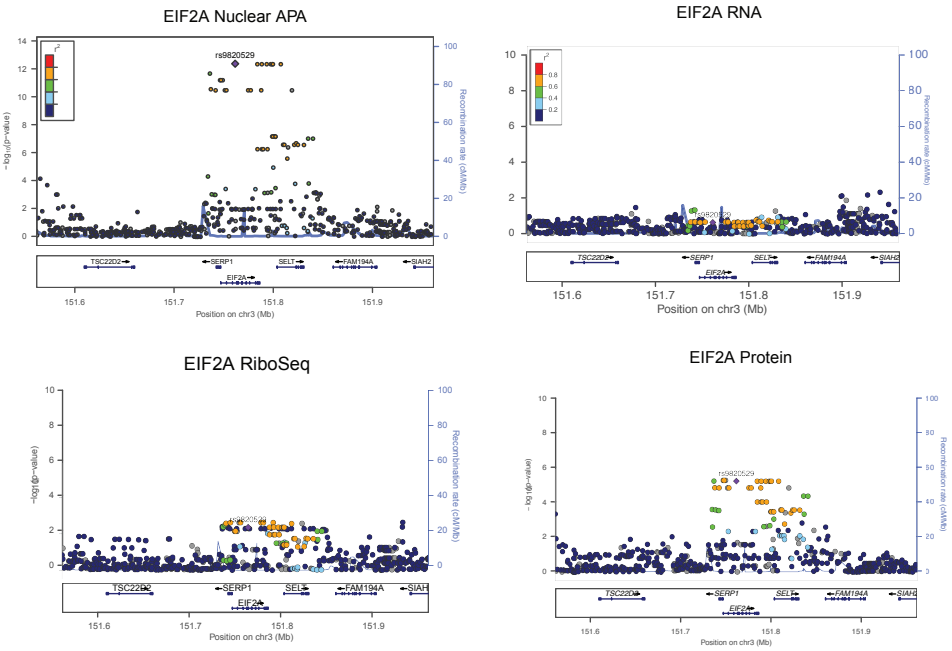

Figure 2.23: **LocusZoom plots for EIF2A molecular associations, Supplement to Figure 2.4B** LocusZoom plots for EIF2A apaQTL in Figure 4b along with associations with RNA expression, ribosome occupancy (ribo-seq), and protein expression as determined using normalized data from Li et al. 2016 [44]. LD patterns were colored according to the HapMap YRI lines.

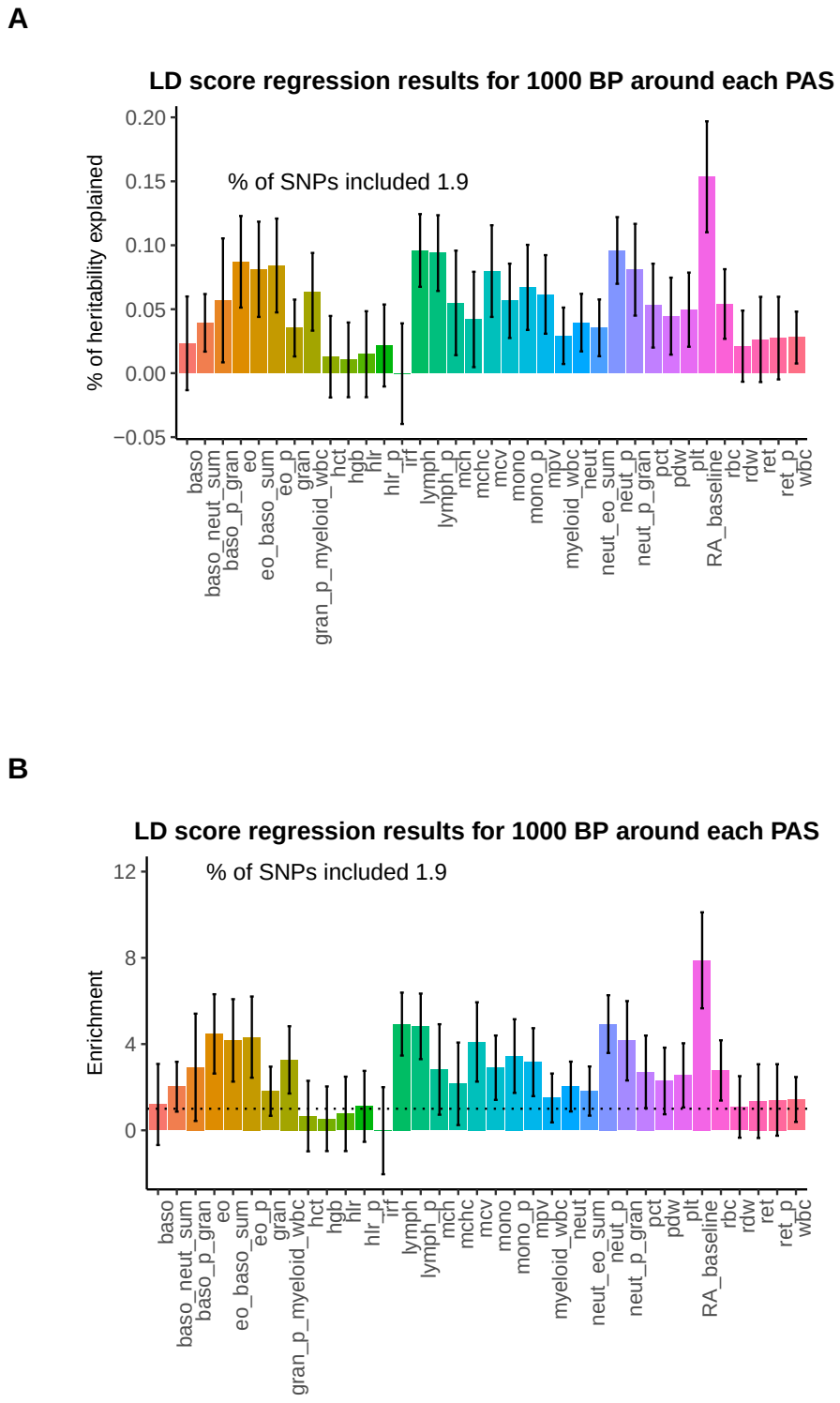

Figure 2.24: **LD Score regression enrichment estimates suggest that APA regulation is likely relevant for complex human phenotypes (A)** Percent of heritability explained by SNPs within 1kbp around each PAS.

Figure 2.24: (continued) Error bars represent +/- 1 standard deviation.Blood phenotype statistics published in Astle et el [2]. Rheumatoid arthritis statistics were obtained from Okada et al [64]. **(B)** Enrichment of heritability explained by SNPs within 1kbp around PAS for the phenotypes analyzed.

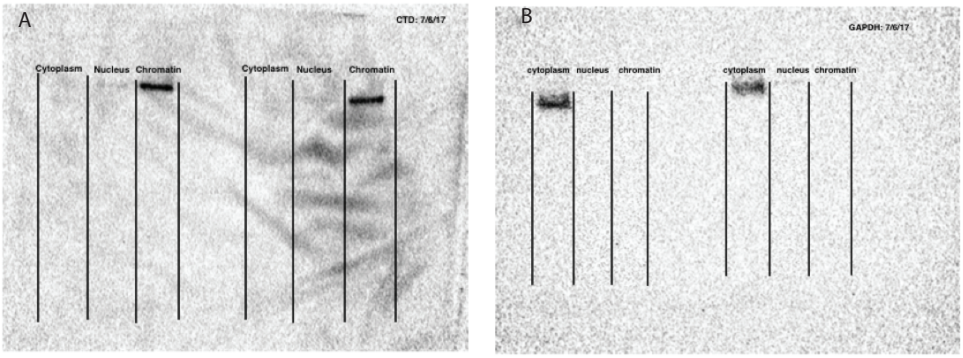

Figure 2.25: **Western Blots to demonstrate cell fractionation (A)** Western blot against Carboxyl terminal domain of RNA Polymerase II, photo captured at 10 second exposure. Blot is not used for quantification, but to validate cell fractionation. **(B)** Western blot against GAPDH to mark glycolysis 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 [53]

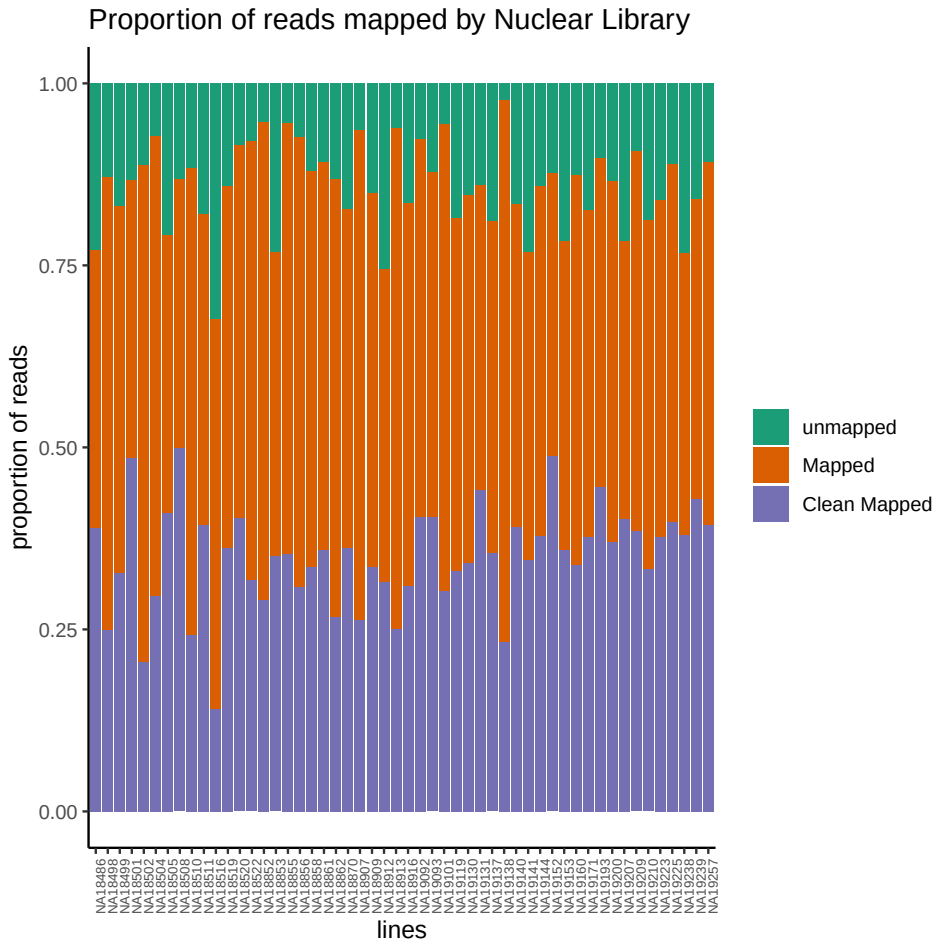

Figure 2.26: **3'-Seq read mapping proportions for the nuclear mRNA fraction** Proportion of reads that map to the genome (mapped) and the proportion of final reads used for analysis are cleanly mapped (Clean Mapped) by nuclear mRNA library. Cleanly mapped reads are reads that mapped successfully and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

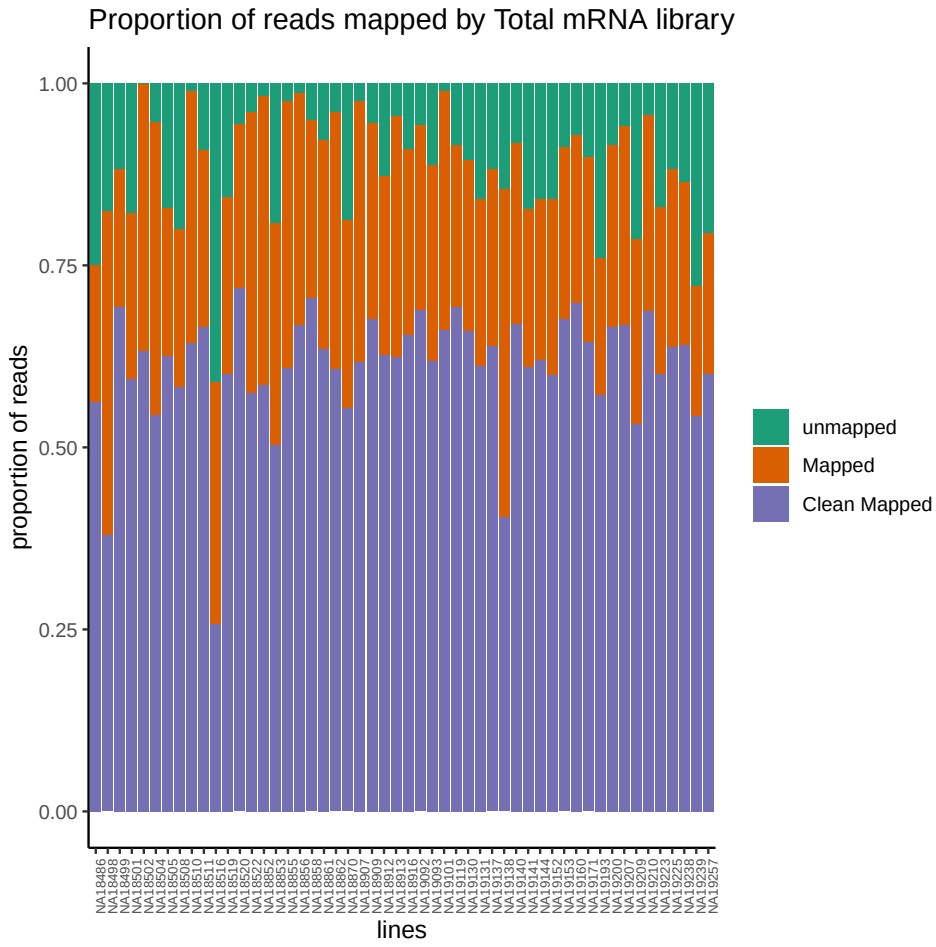

Figure 2.27: **3'-Seq read mapping proportions for the total mRNA fraction** Proportion of reads that map to the genome (mapped) and the proportion of final reads used for analysis that are cleanly mapped (Clean Mapped) by total mRNA library. Cleanly mapped reads are reads that mapped successfully and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

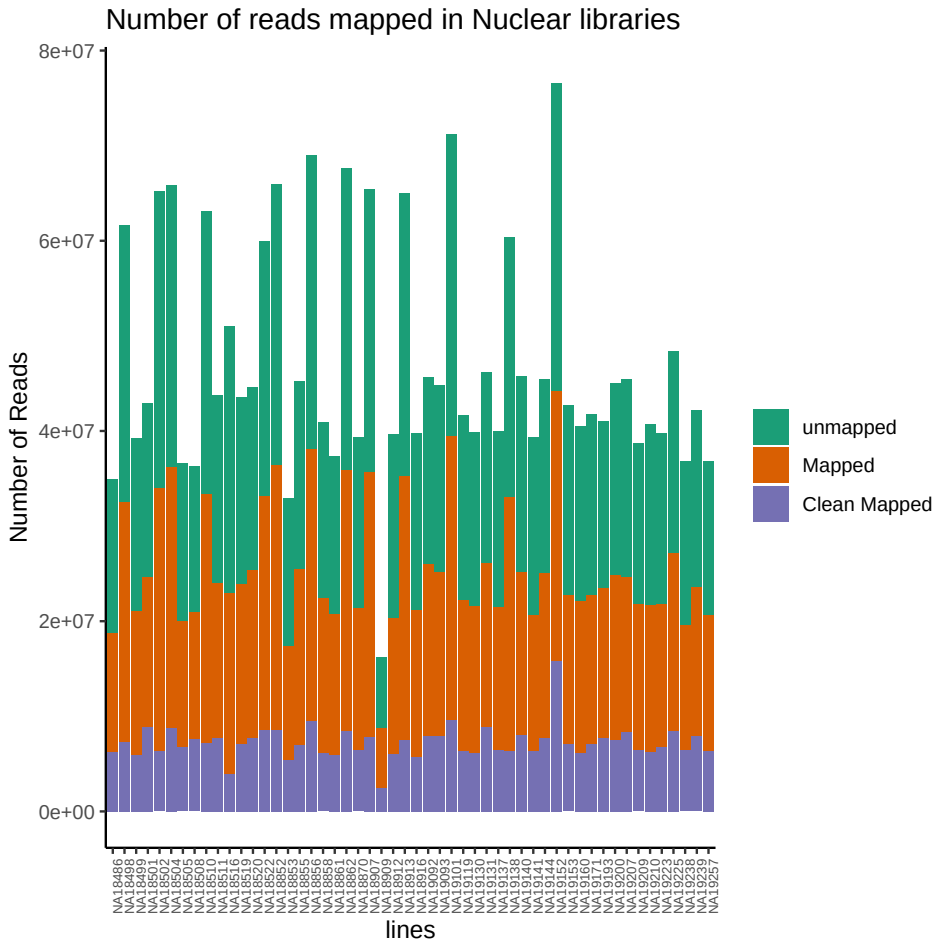

Figure 2.28: **3'-Seq reads mapping counts for the nuclear mRNA fraction** Total number of reads that map to the genome (mapped) and the number of final reads used for analysis that are cleanly mapped (Clean Mapped) by nuclear mRNA library. Cleanly mapped reads are reads that mapped successfully and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

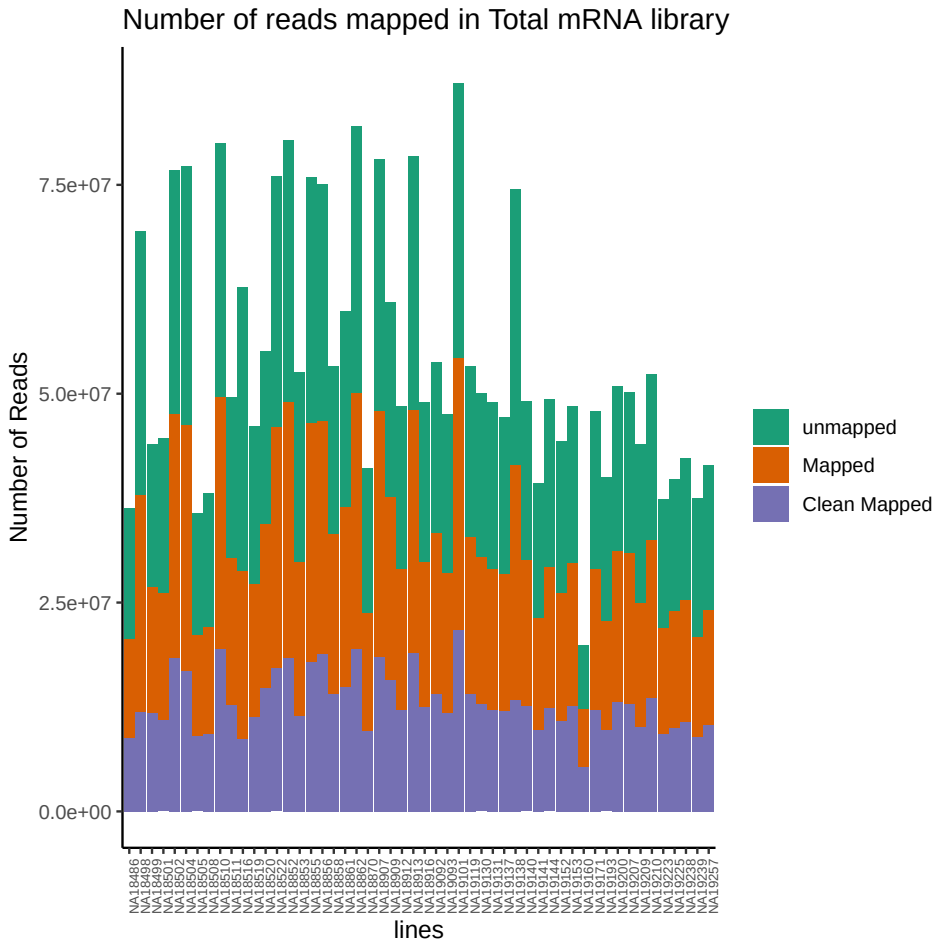

Figure 2.29: **3'-Seq reads mapping counts for the total mRNA fraction** Total number of reads that map to the genome (mapped) and the number of final reads used for analysis that are cleanly mapped (Clean Mapped) by total mRNA library. Cleanly mapped reads are reads that mapped successfully and passed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

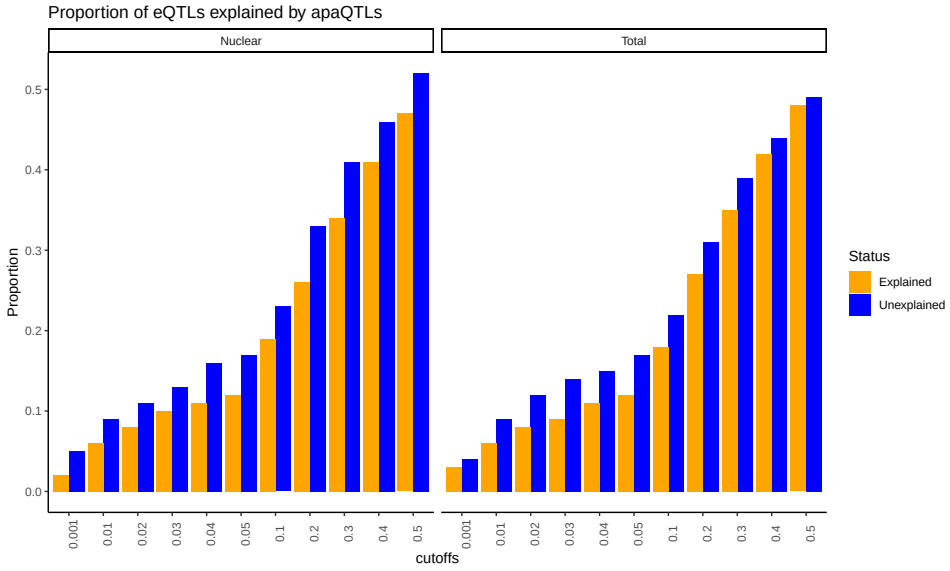

Figure 2.30: **Proportion of eQTLs explained by apaQTLs** Proportion of eQTLs putatively explained by apaQTLs separated by fraction. Expression QTLs could be explained by apaQTLs identified from both fractions. This observation is robust to apaQTL association p-value cutoffs. We observed that apaQTLs explain a slightly higher proportion of previously unexplained eQTLs. Explained/Unexplained status of each eQTL was determined previously in Li et al. 2016 [44]

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 [68] and Li et al. 2016 [44], 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 [44].

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

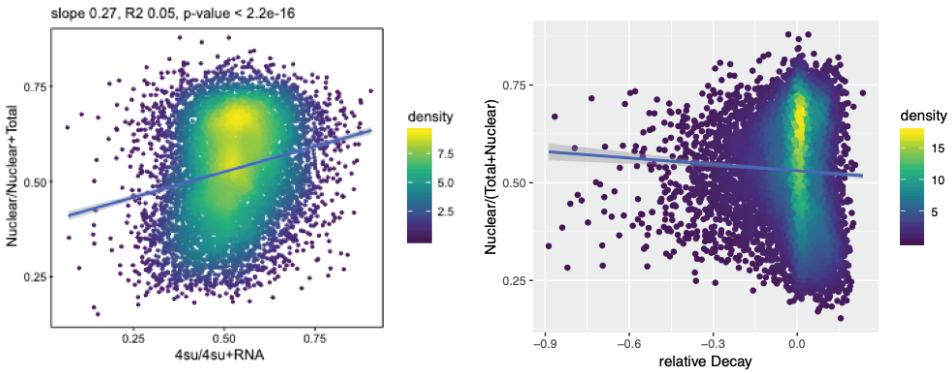

Figure 2.31: **Relationship between 3' Seq and nascent transcription (A)**Nuclear 3' Seq captures polyadenylation of nascent transcripts. The ratio of new mRNA to steady-state mRNA (x axis) are plotted against the ratio of 3' Seq reads from the nuclear fraction to 3' Seq reads from the total mRNA fraction (y axis). Slope, R2, 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, R2, 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 [45]. 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 [16]. 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.

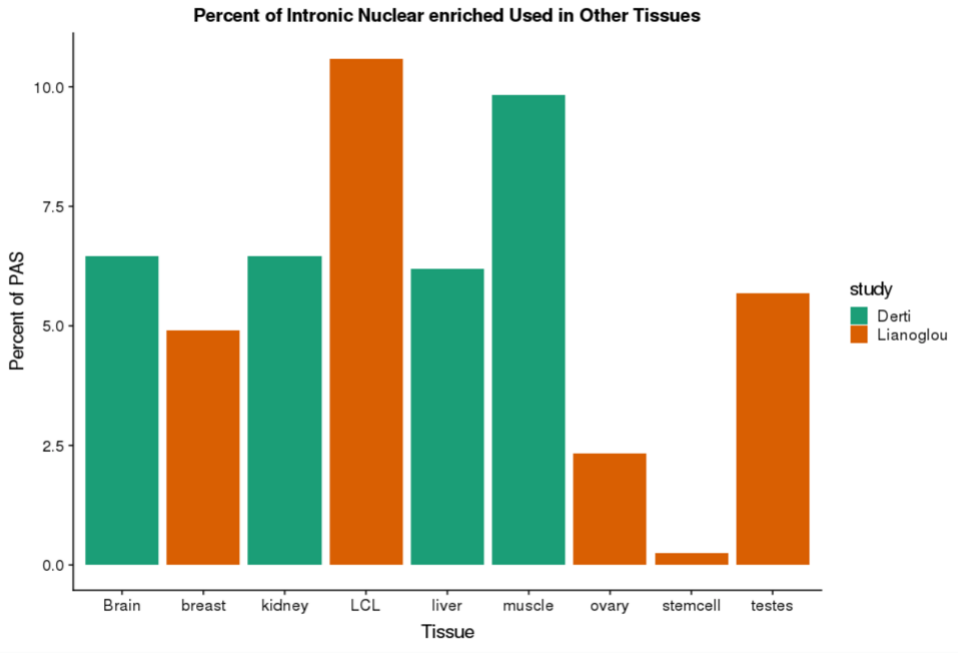

Figure 2.32: **Intronic PAS Discovered in other tissues** Intronic PAS enriched in the nuclear mRNA fraction of LCLs as detected in the total mRNA fraction of other human tissues. Barplot showing the percent of nuclear intronic PAS (of 387) discovered in whole cell 3' Seq from Derti et al. [16], or Lianoglou et al. [45] 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

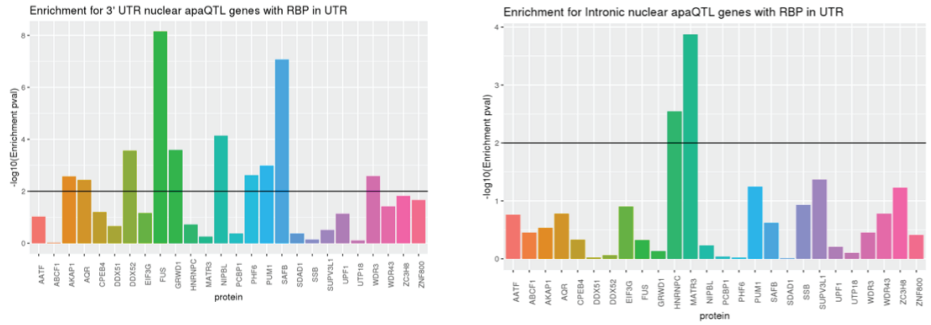

Figure 2.33: **Enrichment for RNA binding in K652 cells(A)**Enrichment for K562 cell RBP binding in 3' UTRs of genes with apaQTLs most strongly associated with a PAS in 3' UTRs compared to genes without apaQTL **(B)** Enrichment for K562 cell RBP binding in 3' UTRs of genes with apaQTLs most strongly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

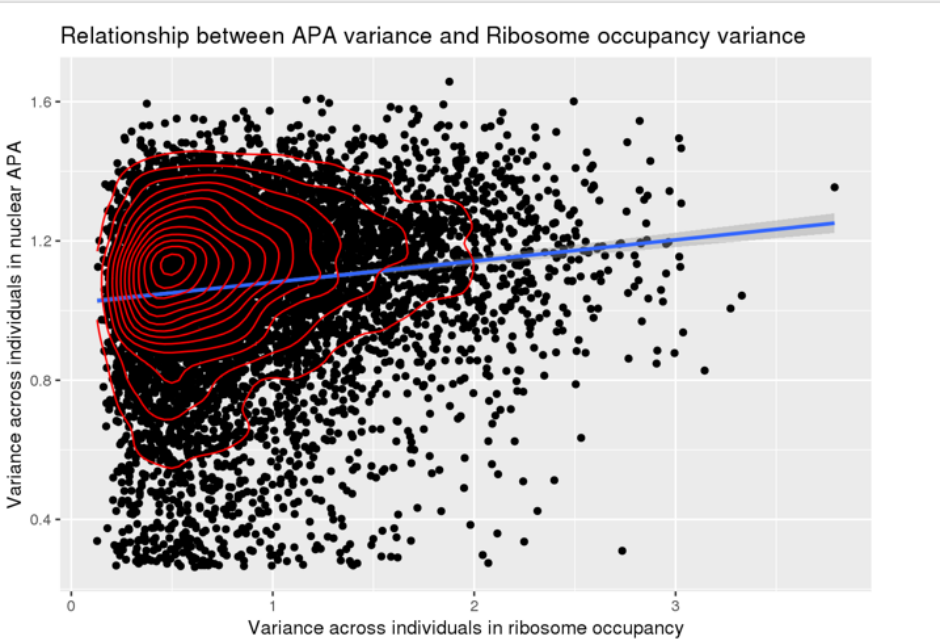

Figure 2.34: **Variance in APA and Ribosome Occupancy** Individual usage variance of the most highly used PAS in each gene (x axis) correlates with individual variance in ribosome occupancy (y axis) as measured in Li et al 2016.[44]

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

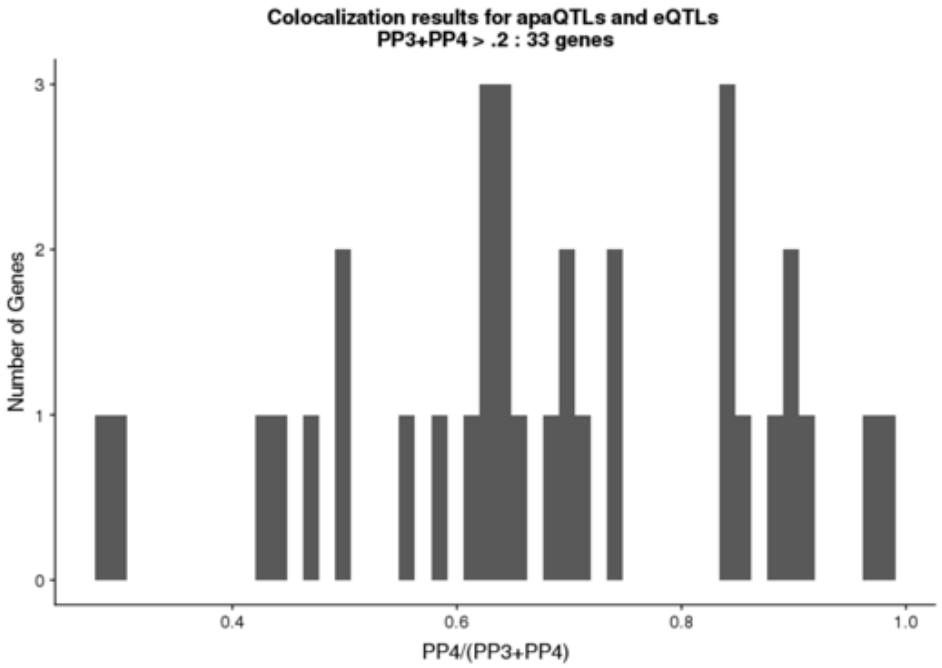

Figure 2.35: **Colocalization of apaQTLs and eQTLs** The apaQTL and eQTLs for the large majority of genes that have both are more likely to colocalize than not. Histogram of number of genes with an apaQTL and eQTL for different values of PP4/(PP3 +PP4).

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 be

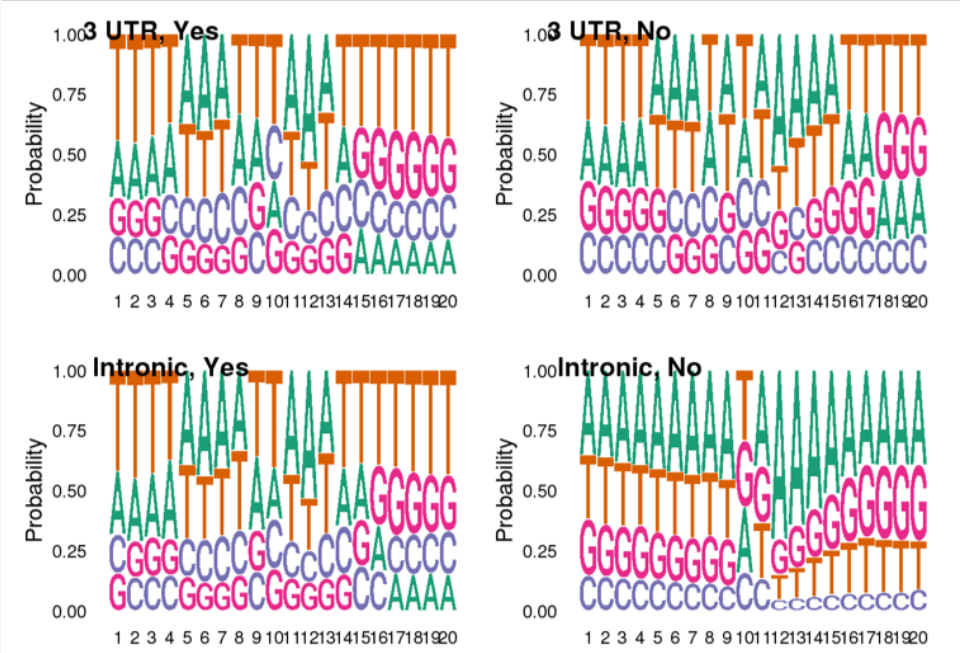

Figure 2.36: **Base Composition around PAS** Position weight matrices representing base composition 10 bps upstream and downstream of identified PAS separated by location and presence/absence of site in polyA DB.

some amount of mispriming for intronic PAS that are not annotated in the polyA DB. 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

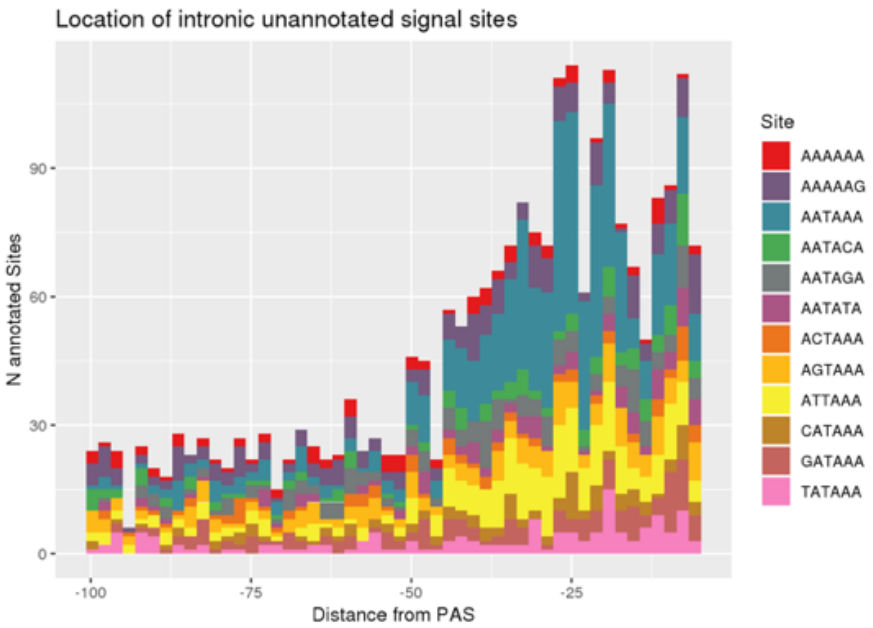

Figure 2.37: **Signal site distribution for intronic unannotated PAS** Stacked histogram of polyadenylation signal sites upstream of unannotated intronic PAS. Distribution similar in shape and structure to that in Figure 2.1D.

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.

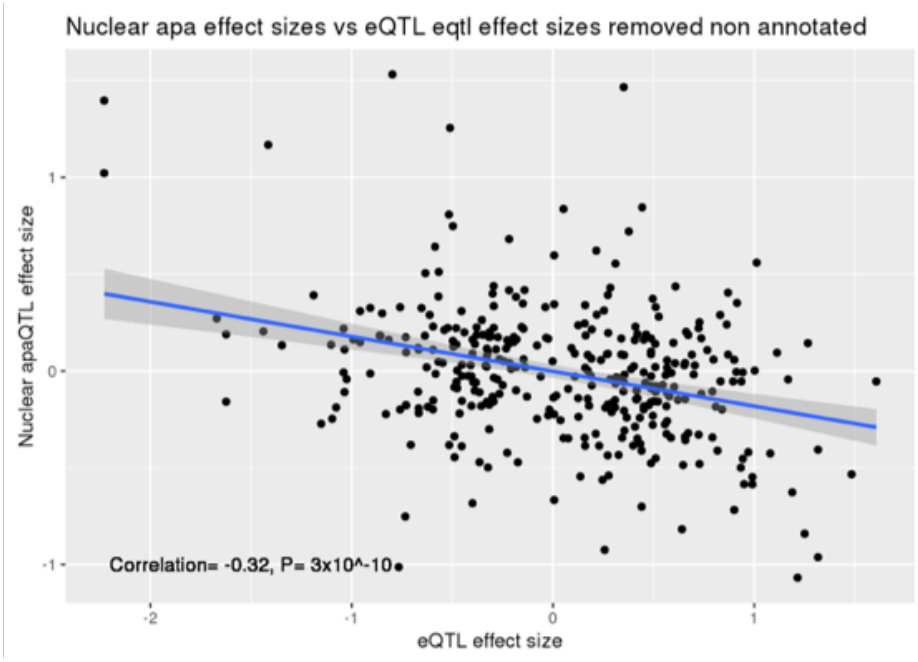

Figure 2.38: **Figure 2.3A without unannotated intronic PAS** Scatter plot of intronic apaQTL effect sizes after removing associations with unannotated intronic PAS plotted against their eQTL effect sizes. Supplemental to Figure 2.3A.

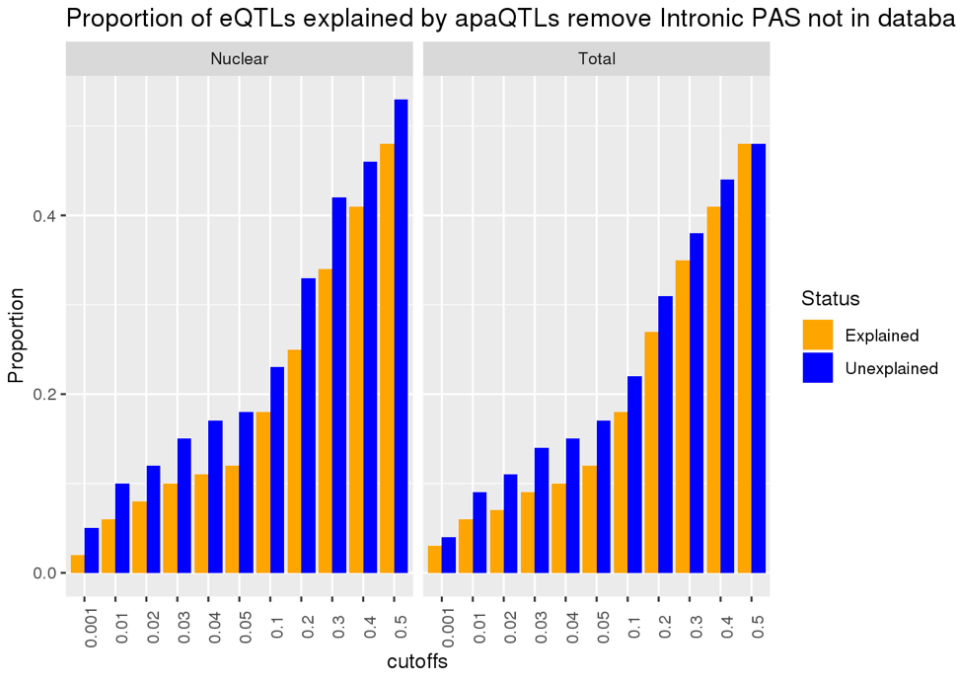

Figure 2.39: **Proportion eQTL explained without unannotated intronic PAS** Proportion of putatively explained by apaQTLs separated by fraction after removing associations with unannotated intronic PAS. Expression QTLs could be explained by apaQTLs identified from both fractions. This observation is 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.[44]

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[44].

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

PREDICTING SUSCEPTIBILITY TO TUBERCULOSIS BASED ON GENE EXPRESSION PROFILING

3.1 Abstract[1]

3.2 Introduction

3.3 Results

3.4 Discussion

3.5 Methods

1. Citation for chapter: John D Blischak*, Ludovic Tailleux*, Marsha Myrthil, Luis B Barreiro, and Yoav Gilad. Predicting susceptibility to tuberculosis based on gene expression profiling. In preparation. * denotes equal contribution.

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 (PolII) elongation speed variation genome wide in a panel of YRI LCLs. Our goal was to map genetic variation associated with PolII 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 PolII 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 [44, 59, 15, 21, 3]. 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 Li et al. quantified the proportion of 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 [44]. 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 [44, 60]. 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 (PolII) density, researchers have discovered that PolII moves along gene bodies at a non-uniform rate [54, 62, 51, 14, 31]. Specifically, PolII density increases proximal to the promoter, at intron exon boundaries, and

at the transcription end site (TES) suggesting PolII pauses at each of these locations during transcription [1, 94, 72]. According to studies in human and other model systems, PolII pausing is tightly regulated [62, 9, 71, 24]. While, various studies have mechanistically implicated PolII dynamics in alternative splicing and APA, there is still debate surrounding causal relationships and the degree to which PolII pauses or simply slows down [70, 73]. Moreover, despite the large body of molecular work, no study has quantified interindividual variation in PolII elongation rate.

We suspect genetic variation contributing to differences in PolII 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 PolII density as an estimate of PolII elongation rates. We intended to map genetic variation associated with elongation differences to ask if PolII 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 PolII density and nascent transcription genome wide [88]. For this project, we decided to use the Nascent Elongating Sequencing (NET-seq) protocol published by Andreas Mayer and L. Stirling Churchill in 2016 [53]. The protocol maps PolII 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 ([54], 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 PolII density and map

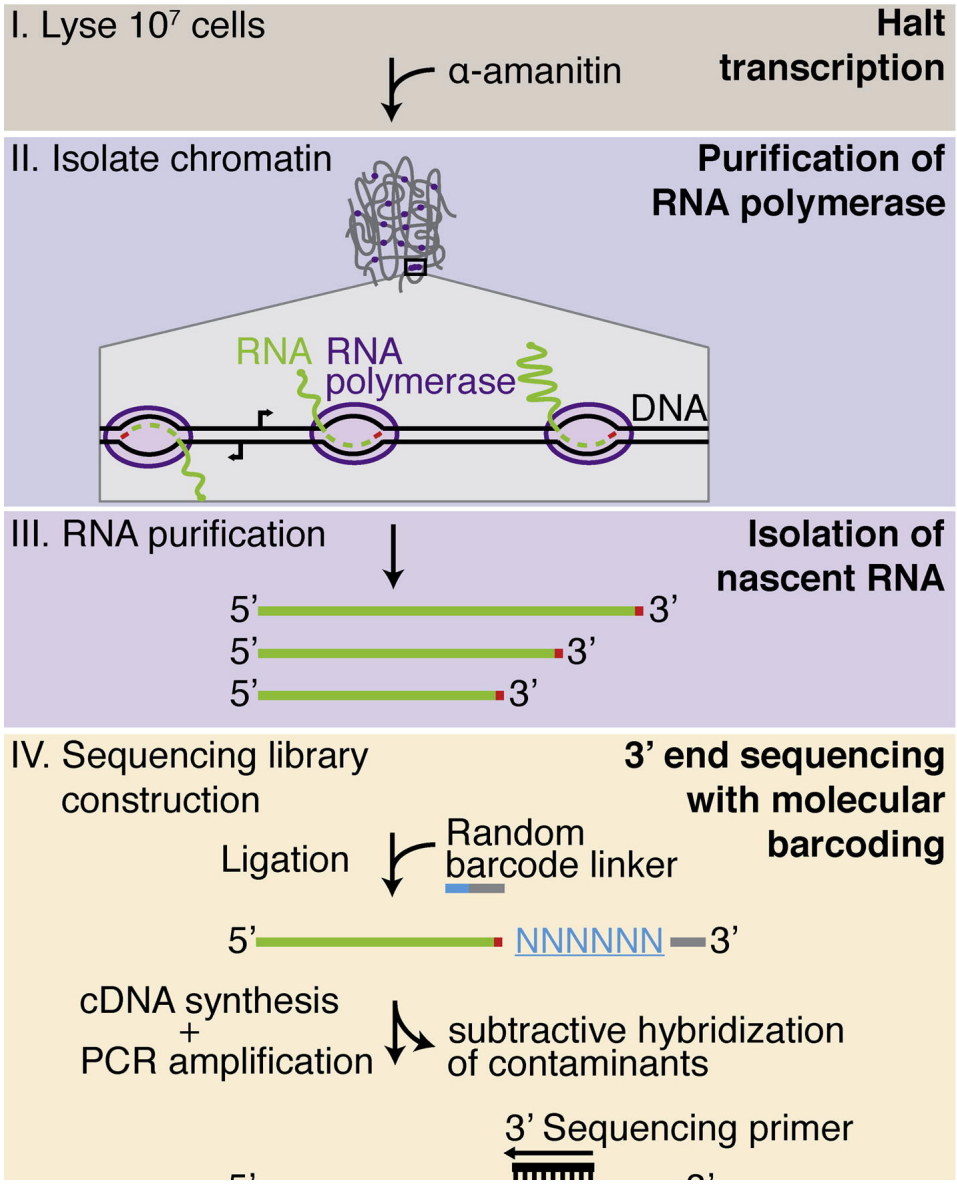

Figure 4.1: **Graphical representation of NET-seq protocol published in Mayer et al. [54] Panel I:** Halt transcription in cells with α-Amantin. **Panel II:** Purify chromatin containing PolII. **Panel III:** Purify nascent mRNA from chromatin fractionation. **Panel IV:** Library construction by adding DNA linker to 3' OH group, cDNA synthesis, and removal of mature mRNA contaminates

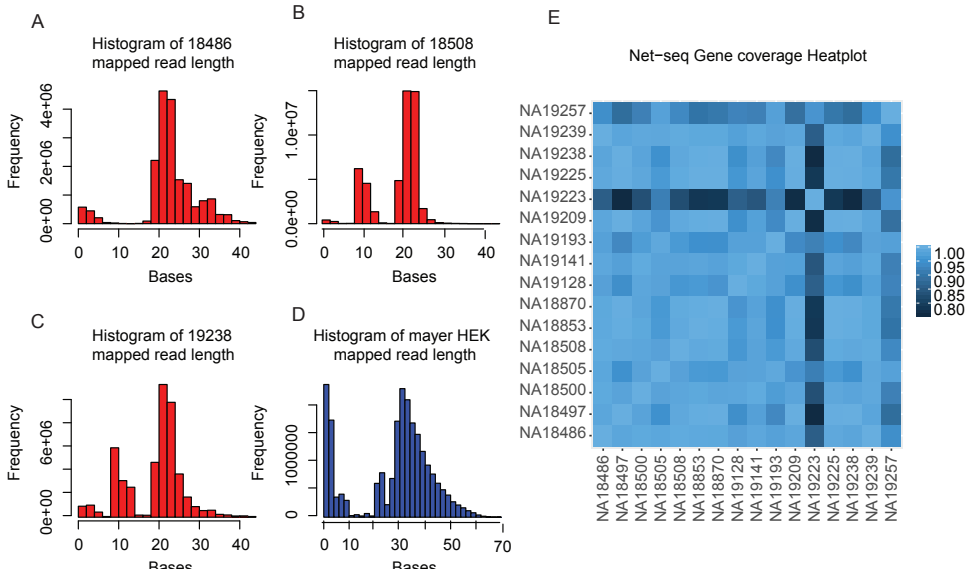

Figure 4.2: **Quality control metrics for NET-seq libraries. A** Histogram of mapped read lengths for NET-seq library NA18486. **B** Histogram of mapped read lengths for NET-seq library NA18508. **C** Histogram of mapped read lengths for NET-seq library NA19238. **D** Histogram of mapped read lengths for NET-seq library generated from HEK cells and published in Mayer et al [54]). **E** Pearson correlations for NET-seq library coverage in gene bodies. Coverage calculated as number of reads mapping to gene standardized by gene length.

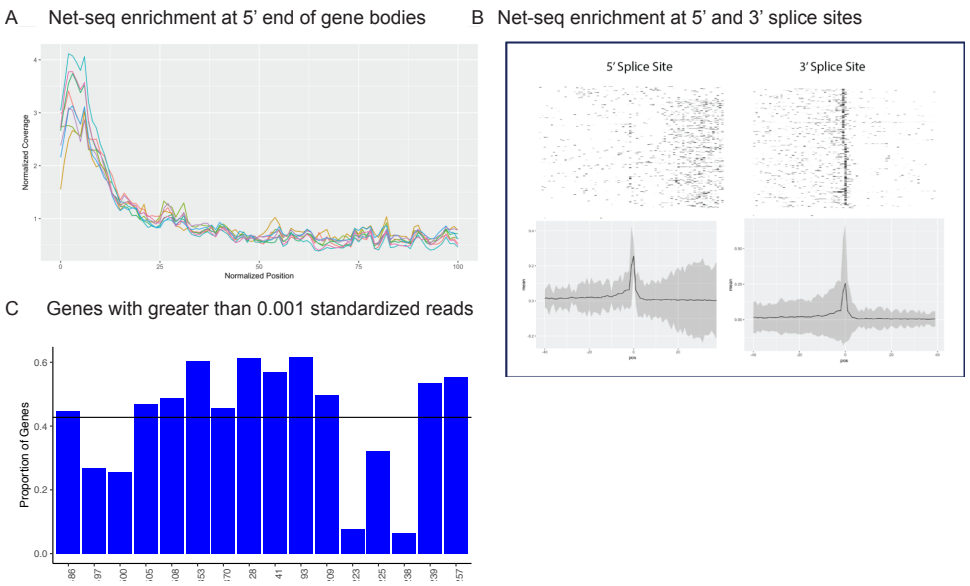

Figure 4.3: **NET-seq Gene coverage.** **A** Enrichment for Distribution of read coverage along gene bodies calculated with Picard tools (NA18505, NA18508, NA18486, NA19239, NA19239, NA19141, NA19193, NA19257, NA19128). **B**Histogram and smoothed density plots for NET-seq read coverage at 5' and 3' 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 PolII 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 ([89], methods). For ACTB, the smoothing allowed us to identify regions of likely Pol II pausing at the 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 regions may be mipmapping, therefore contributing to regions of high read density.

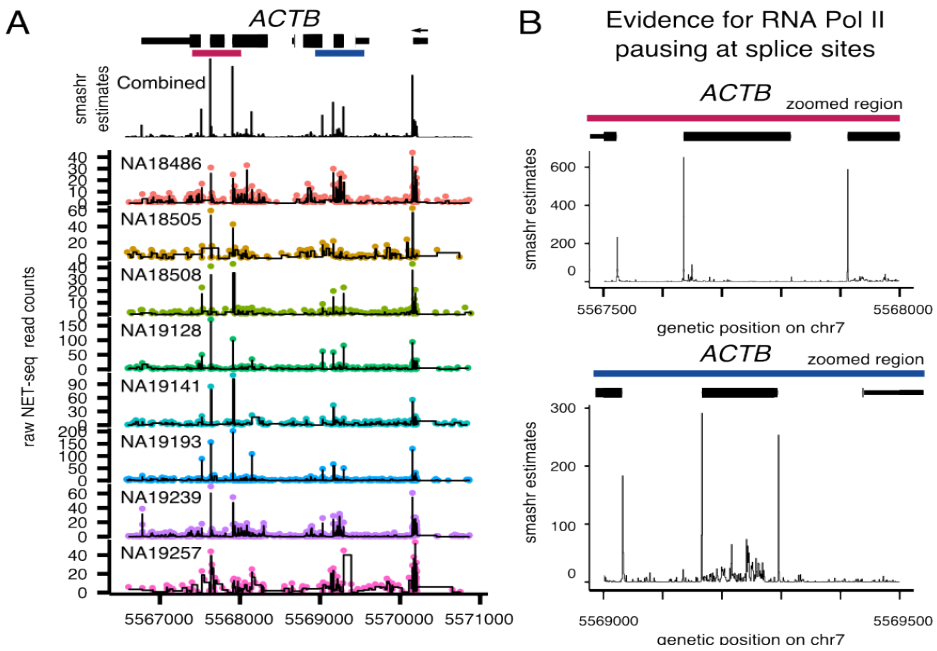

Figure 4.4: **Smoothing of NET-seq data using smashr [89]. A** Raw counts along ACTB gene locus for 8 combined libraries (NA18505, NA18508, NA18486, NA19239, NA19239, NA19141, NA19193, NA19257, NA19128) as well as each library separately. **B** Smoothed coverage along ACTB on combined NET-seq data. Signal at 5' and 3' splice sites.

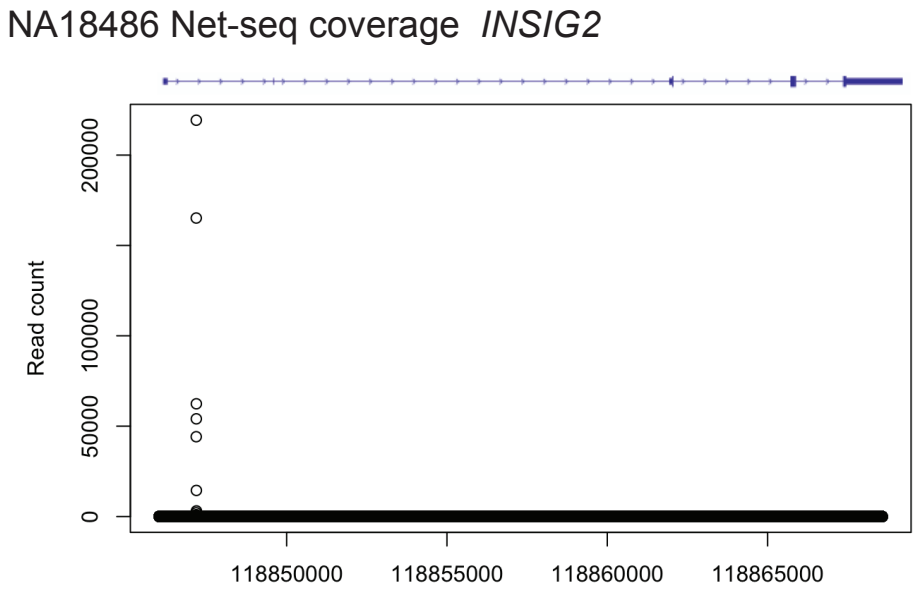

Figure 4.5: **NA18486 NET-seq coverage along INSIG2 locus**

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 (PolII) 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 [54]. 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 regions of high PolII 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 lot 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 [54]. 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 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 [54]. 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) [54]. 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 PolII patterns. Unfortunately, our low coverage libraries were too noisy to confidently quantify PolII density, even if we merged on genotypes.

We still believe that mapping genetic variants associated with PolII density would likely help in our understanding on gene regulation. There are a variety of alternative methods to collect nascent mRNA or PolII associated transcripts. Other potential methods largely fall into two classes. The first category relies on immunoprecipitation of PolII or a specific post translationally modified version of PolII [62, 23, 7, 35, 88, 55]. These methods potentially suffer from non-specific binding of anti-bodies, and with the exception of mNetseq, are restricted to 200bp resolution [62]. 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 [38, 13, 83, 22, 51, 55, 88].

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 [37]. 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 [37]. It is possible that once published, we or others could use these data to test for genetic variation associated with co-transcriptional PolII 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 [28]. 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. [60].

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 [60]. We used the Native Elongating Transcript sequencing (NET-seq) collection protocol published in Mayer and Churchman 2016 with minor adjustments to three buffers [56]. 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 [56]. 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 [56].

4.5.3 Data processing

We mapped all NET-seq libraries to GRCH37.75 downloaded from Ensemble using subjunc with default settings [29, 46] We used umi_tools extract to extract the 6 base UMIs and umi-tools dedup to collapse duplicate reads [78].

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 [19**?**]. We used featureCounts with the -T 5 flag to quantify reads within the gencode.v19 gene annotation [19**?**]. We used pysam to extract mapped read statistics from bam files [41]. We downloaded the HEK NET-seq published in Mayer et al. available from GEO under accession number GSE61332 [53]. 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 [89]. The

function implemented a wavelet-based Empirical Bayes shrinkage method [89].

We attempted to recreate a version of Mayer et al. figure 7 using NA18486 sequence coverage for this analysis [53]. 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

5.1 A joint Bayesian model provides a general framework for analyzing functional genomics studies with many conditions

5.2 Initial success classifying individuals susceptible to tuberculsosis and future directions

5.3 Incorporating lessons from single cell pilot study for future studies of the genetic basis of gene expression noise and the response to bacterial infection

5.4 The importance of mitigating batch effects in any genomics experiment

5.5 Concluding remarks

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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

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

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

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

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

[72] 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.

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

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

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

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

[91] 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.

[92] 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.

[93] 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.

[94] 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.