Alternative polyadenylation mediates genetic regulation of gene expression

Briana Mittleman[1], Sebastian Pott[2], Shane Warland[3], Tony Zeng[3], Mayher Kaur[3], Yoav Gilad[2,3], Yang I Li[2,4]

[1] Genetics, Genomics, and Systems Biology, University of Chicago, IL

[2] Department of Human Genetics, University of Chicago, IL

[3] Section of Genetic Medicine, Department of Medicine, University of Chicago, IL

Correspondence should be addressed to Y.G. (gilad@uchicag.edu) or Y.I.L. (yangili1@uchicago.edu)

Abstract

With the exception of mRNA splicing, little is known about co-transcriptional or post-transcriptional regulatory mechanisms that link noncoding variation to variation in organismal traits. To begin addressing this gap, we used 3' Seq to characterize alternative polyadenylation (APA) in the nuclear and total RNA fractions of 52 HapMap Yoruba lymphoblastoid cell lines, which we have studied extensively in the past. We identified thousands of polyadenylation sites that are differentially detected in nuclear mRNA and whole cell mRNA, and found that APA is an important mediator of genetic effects on gene regulation and complex traits. Specifically, we mapped 602 apaQTLs at 10% FDR, of which 152 were found only in the nuclear fraction. Nuclear-specific apaQTLs are highly enriched in introns and are also often associated with changes in steady-state expression levels, suggesting a widespread mechanism whereby genetic variants decrease mRNA expression levels by increasing usage of intronic PAS. We identified 24 apaQTLs associated with protein expression levels, but not mRNA expression, and found that eQTLs that are not associated with chromatin QTLs are enriched in apaQTLs. These findings support multiple independent pathways through which genetic effects on APA can impact gene regulation. Finally, we found that 19% of apaQTLs were also previously associated with disease. Thus, our work demonstrates that APA links genetic variation to variation in gene expression levels, protein expression levels, and disease risk, and reveals uncharted modes of genetic regulation.

Introduction

Nearly all 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 (Chun et al., 2017) found that the majority of autoimmune GWAS loci do not appear to affect mRNA expression levels. Two other studies observed that many genetic variants affecting variation in protein levels (pQTLs) do not affect mRNA expression levels (Battle et al., 2015; Chick et al., 2016). 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 is known to impact mRNA expression levels, the mechanisms by which expression is affected is often unclear. Indeed, a third of all eQTLs identified in human lymphoblastoid cell lines (LCLs) are not associated with any chromatin-level regulatory phenotypes including transcription factor binding and histone modifications (Y. I. Li et al., 2016), again raising the possibility that understudied regulatory mechanisms mediate these eQTL effects.

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 (Tian & Manley, 2017; Mayr, 2016; Shi, 2012). 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 (Mayr, 2017; Tian & Manley, 2017)), 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 have phenotypic effects. For example, global increased usage of intronic PAS has been shown to increase risk

for multiple myeloma and chronic lymphocytic leukemia (Lee et al., 2018; Singh et al., 2018) by producing mRNAs that translate into truncated proteins, which causes defects in tumour-suppressive functions (Lee et al., 2018; Singh et al., 2018).

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 (L. Li, Gao, Peng, Wagner, & Li, 2019; Yoon, Hsu, Im, & Brem, 2012; Yang et al., 2019; Bonder et al., 2019). While using existing RNA-seq to study APA is economical, identifying PAS and estimating usage using RNA-seq are error-prone and often imprecise (Ha, Blencowe, & Morris, 2018). Furthermore, using existing whole-cell, total RNA-seq data 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 (Degner et al., 2012; McVicker et al., 2013; Y. I. Li et al., 2016; Pickrell et al., 2010). Integrating the apaQTLs we identified with previously collected data types allowed us to characterize the functional impact of variation in APA on each of the major steps of the gene regulatory cascade. We use these data to show that genetic effects on polyadenylation can independently affect virtually all steps of gene regulation (mRNA expression level, translation rate, and protein expression level), and that such effects can be associated with protein expression, but not RNA expression.

Results

To measure the impact of inter-individual variation in APA on multiple stages of gene regulation, we quantified PAS usage in a panel of 52 Yoruba HapMap LCLs. These same samples have been the subjects of multiple studies of gene regulation over the last decade (Degner et al., 2012; McVicker et al., 2013; Y. I. Li et al., 2016; Pickrell et al., 2010). We applied 3′ Seq to mRNA collected from whole cells (total fraction) of 52 LCLs to comprehensively identify PAS and estimate usage without relying on existing annotations. In addition, to capture polyadenylated mRNA that may be under-represented or absent in the total fraction

due to rapid turnover, we separately applied 3′ Seq to mRNA from isolated nuclei (nuclear fraction) of the same 52 LCLs (Fig. 1A).

Nuclear 3′ Seq measures PAS usage independent of post-transcriptional decay

We first verified that the nuclear 3′ Seq data capture transcripts at a more primitive stage compared to the total 3′ Seq data, and thus better reflect mRNA diversity independent of decay. To this end, we reasoned that genes with higher nuclear 3′ Seq read counts relative to total 3′ Seq read counts should show faster decay on average. Indeed, we found that the relative number of nuclear 3′ Seq reads to total 3′ Seq reads is positively correlated with both the ratio of 4sU-seq to RNA-seq read counts (Y. I. Li et al., 2016) ($p < 2.2^{-16}$, Fig. 1B) and direct gene level mRNA decay estimates (Pai et al., 2012) ($p < 2.2^{-16}$, Supplementary Fig. 1), two different measures of decay. After filtering the 3′ Seq data for possible internal priming (Methods), we identified 41,810 PAS in 15,043 genes. 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 (Tian & Manley, 2017; Mayr, 2016; Shi, 2012). We found that the polyA binding protein motif (AATAAA), also known as the polyadenylation signal site, is the most strongly enriched protein binding motif in regions surrounding our PAS ($p < 10^{-391}$). 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%,) (Fig. 1C, Supplementary Fig. 3) and that nearly half (48.3%) of all 41,810 PAS we identified are located in 3′ UTRs (19.4x enrichment) (Singh et al., 2018). 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%) (Fig. 1D, Supplementary Fig. 2). 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, Fig. 1D). Thus, the recognition of intronic polyadenylation signals is a general mechanism that can result in premature termination of transcription. In addition, although slightly enriched in the first introns of genes (2.69x enrichment over uniform distribution), intronic PAS could be identified in all introns and thus are not likely to result from defective telescripting activity alone (e.g. from a depletion in U1 small nuclear ribonucleoprotein (snRNP)) (Kaida et al., 2010; Berg et al., 2012; Oh et al., 2017).

We also observed that intronic PAS have on average lower usage across individuals than PAS located in 3′ UTRs (16.9% vs 46.2%). These differences may be explained by weaker polyadenylation signals at intronic PAS compared to 3′ UTR PAS, but we hypothesized that some intronic PAS might have low usage because

premature polyadenylation at intronic sites can produce short-lived transcripts that are rapidly degraded and thus 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 fraction. By comparing PAS usage estimated in the nuclear and total fractions from all 52 individuals, we identified 591 PAS in 585 genes that are used more often in the nuclear fraction (10% FDR). 134 of these 591 PAS were found to be used by 1% or less of the transcripts in the total fraction, suggesting that these transcripts may be absent from the cytoplasm (Fig. 1E, Supplementary Fig. 4, Methods). Notably, we found that 387 of the nuclear-enriched PAS are intronic (Supplementary Fig. 4; see example in Fig. 1F), 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 Fig. 5) (Wang, Nambiar, Zheng, & Tian, 2018). These findings suggest that mRNA transcripts are polyadenylated in introns at a higher frequency than generally appreciated, and that many of these isoforms escape detection from studies of total mRNA owing to their rapid decay.

Genetic loci associated with variation in APA

We next sought to identify genetic loci associated with inter-individual variation in APA. We quantified APA as the normalized ratio of reads mapping to each PAS compared to reads mapping to all PAS assigned to the same gene (Methods, Supplementary Fig. 6-7). We tested cis-association between genetic variants and PAS usage, correcting for batch and the top principal components (Methods, Supplementary Fig. 8). Using 3′ Seq data from the nuclear fraction, we identified 602 nuclear apaQTLs in 479 genes at 10% FDR. In the total fraction, we identified 443 apaQTLs in 353 genes at 10% FDR. For example, individuals with the C/C genotype (rs11032578) are more likely to use an intronic PAS in the *ABTB2* gene compared to individuals that are heterozygous C/T or homozygous T/T (Fig. 2A). In both fractions, apaQTLs occur near the PAS they most strongly correlate with and are located at the 3′ ends of gene bodies (Fig. 2B-C). While the proximity of the apaQTLs to PAS may suggest that genetic variants that affect polyadenylation signal motifs drive most of the genetic effects on APA, we found limited evidence that supports this possibility (Supplementary Fig. 9).

Next, we quantified the sharing and specificity of genetic effects on APA in the nuclear and total fractions. We estimated that the vast majority of nuclear apaQTLs were shared with total apaQTLs (π_1 1=0.85) and vice-versa (π_1=0.87, Supplementary Fig. 10). Additionally, we observed that their effect sizes were highly correlated (r^2 = 0.66; $p = 10^{-16}$, Fig. 2D, Supplementary Fig. 11). These results suggest that the predominant mechanism by which genetic variants affect steady-state PAS usage is to directly impact PAS

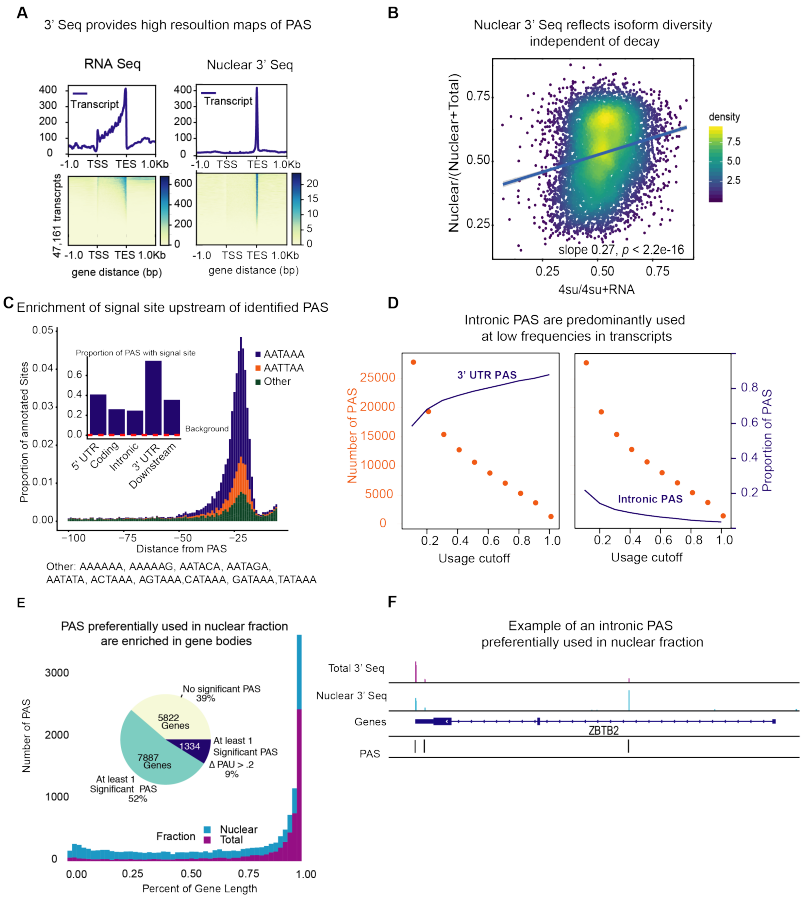

Figure 1: **(A)** Meta gene plot showing read coverage for five RNA sequencing libraries collected from LCLs (Pickrell) *(Left)* and for five 3′ Seq libraries collected from nuclei isolated from LCLs *(Right)*. **(B)** Nuclear 3′ Seq capture polyadenylation of nascent transcripts. The ratio of new mRNA to steady-state mRNA (Y. I. Li et al., 2016) (x axis) are plotted against the ratio of 3′ Seq reads from the nuclear fraction to 3′ Seq reads from the total fraction (y axis). **(C)** *(Main)* Density of canonical (AATAAA, AATTAA) and other polyadenylation signal sites upstream of identified PAS. *(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)** Number (orange) and proportion (purple) of PAS in the 3′ UTR (*Left*) and introns (*Right*) are plotted against usage cutoff in the nuclear fraction. The proportion of intronic PAS increases as the usage cutoff decreases, implying that a disproportionate number of intronic PAS are used at low frequencies. **(E)** *(Main)* Meta gene plot showing the number of differentially used PAS identified by LeafCutter (Online Methods) with a Δ PAU of 0.20 across the gene body. *(Inset)* Estimated number of genes identified with differential PAS usage between total and nuclear fractions. **(F)** *ZBTB2* was identified to harbor a differentially used PAS between total and nuclear fractions. 3′ Seq tracks represent aggregated read counts from all 52 individuals.

choice rather than to affect the stability of an isoform ending at one site relative to that with another ending (e.g. by affecting isoform-specific decay). Nevertheless, we identified 153 nuclear-specific apaQTLs and 97 total-specific apaQTLs (Methods).

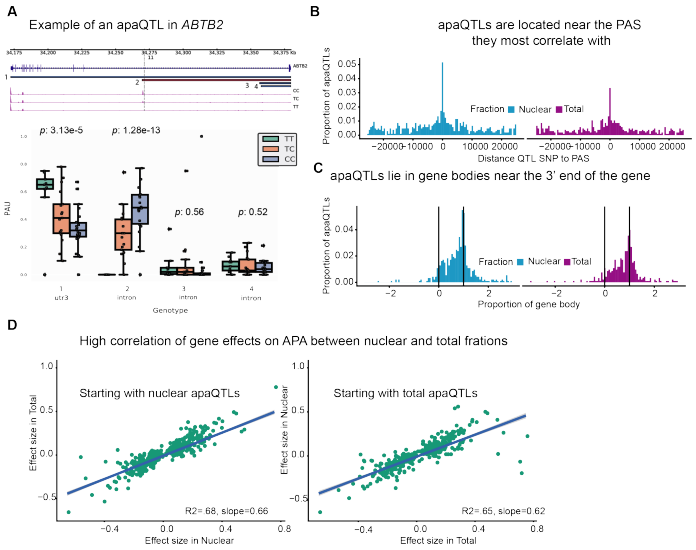

Figure 2: **(A)** An apaQTL in the *ABTB2* gene impact usage of an intronic PAS. *(Top)* Gene track and identified PAS. Each bar represents a potential isoform. The red bar corresponds to the isoform most strongly associated with the apaQTL. The vertical dotted line represents the position of the strongest apaQTL SNP. *(Bottom)* Polyadenylation site usage at each PAS by genotype listed according to isoform order above. The C allele increases usage of the intronic PAS. **(B)** Location of the top nuclear (*Left*) and total (*Right*) apaQTL SNPs relative to their corresponding PAS. **(C)** Meta gene plot showing the distribution of apaQTL SNPs in the annotated gene body, where 0 represents the TSS and 1 represents the annotated transcription end site. **(D)** Effect sizes of apaQTLs originally identified at 10% FDR in nuclear *(Left)* and total *(Right)* fraction plotted against the effect sizes ascertained in total and nuclear fractions, respectively.

APA explain eQTLs that are not associated with chromatin phenotypes

Given that most apaQTLs identified in our study are represented in the nuclear fraction, we focused on the 602 nuclear apaQTLs for subsequent analyses. Although APA produces isoforms with distinct ends, it is possible that the isoforms are functionally identical, especially when they differ only in 3′ UTR lengths and thus encode the same protein sequence. To better understand the functional impact of apaQTLs, we asked whether they are also associated with changes in gene or protein expression levels. We found that genes associated with apaQTLs are enriched for both genes with eQTLs (eGenes, Wilcoxon rank sum test $p = 1.36 \times 10^{-12}$) and genes with protein expression QTLs (pGenes; $p = 0.0006$) compared to genes with neither association (Fig. 3A, Supplementary Fig. 12) (Y. I. Li et al., 2016; Battle et al., 2015). Notably, we found that nuclear-specific apaQTLs are even more enriched for eGenes ($p = 0.002$) compared to apaQTLs that are shared in both fractions. This observation led us to hypothesize that intronic apaQTLs affect gene expression levels by increasing the number of transcripts that use premature intronic PAS, of which many may be subject to rapid decay. Indeed, we found a negative correlation between the genetic effect sizes for intronic PAS usage and mRNA expression levels ($p = 8.97 \times 10^{-7}$, Fig. 3B, Supplementary Fig. 13). Thus,

our analyses suggest a widespread mechanism whereby genetic variants decrease mRNA expression levels by increasing usage of premature PAS located in introns. Of note, 13 of the apaQTLs that were detected only in the nuclear fraction are also eQTLs, which highlights the importance of considering multiple stages of mRNA biogenesis to uncover eQTL mechanisms.

To further investigate the contribution of APA to gene expression, we focused on a set of eQTLs that we previously classified as those with explained putative mechanisms eQTLs (1164 eQTLs, $\sim$ 60%) or as unexplained eQTLs (801 eQTLs, $\sim$ 40%) using data from the same LCLs (Y. I. Li et al., 2016). 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). By contrast, 40% of eQTLs were not associated with any chromatin-level measures, and thus their mechanisms of action remain unknown. 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 ($p = 0.01$) (Fig. 3C, Supplementary Fig. 14). We also found that apaQTLs exhibited a chromatin enrichment profile that was more similar to the unexplained eQTLs than the explained eQTLs. In particular, apaQTLs and unexplained eQTLs were more likely to lie in regions of transcription elongation or are associated with weak transcription, and less likely to lie in enhancers or promoters than explained eQTLs (Fig. 3D. Overall, we estimated that the apaQTLs can provide a putative mechanism for 17.3% of otherwise unexplained eQTLs (see Methods). For example, an unexplained eQTL for *C10orf88* (rs7904973) colocalizes with an apaQTL associated with increased usage of an intronic PAS (Fig. 4A). This observation thus highlights APA as one important mechanism by which genetic variation impacts gene expression without affecting enhancer and promoter activity.

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 (Y. I. Li et al., 2016; Battle et al., 2015). However, we found 24 apaQTLs that affect protein expression, but not mRNA expression (Supplementary Table 1), suggesting a more complex mode of gene regulation independent of mRNA expression that involves APA. We found that five of these 24 apaQTLs were significantly associated with ribosome occupancy (Supplementary Table 1). This finding

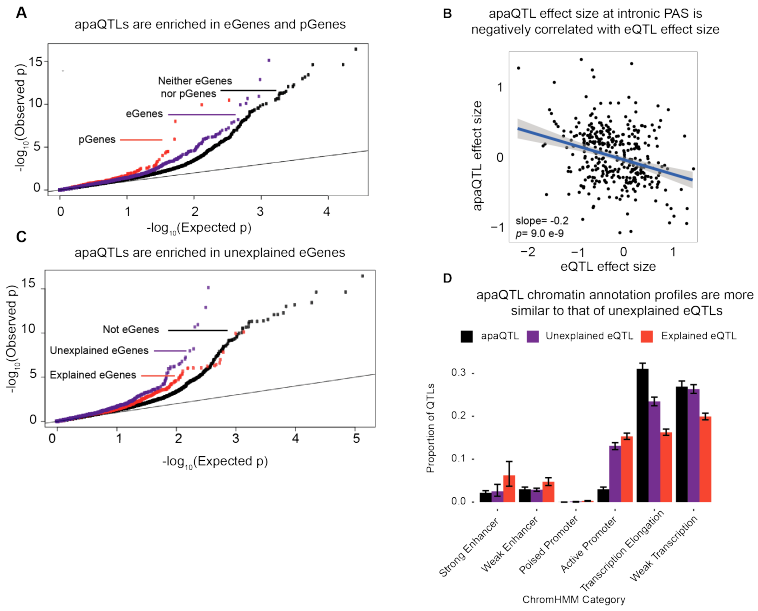

Figure 3: **(A)** Quantile-quantile (Q-Q) plot for apaQTLs shows an enrichment in both eGenes and pGenes. **(B)** Scatter plot of intronic apaQTL effect sizes plotted against their eQTL effect sizes shows negative correlation. **(C)** Quantile-quantile (Q-Q) plot for apaQTLs shows that apaQTLs are more highly enriched in unexplained eGenes compared to explained eGenes. **(D)** Proportion of apaQTL, explained eQTL, and unexplained eQTL SNPs in different genomic annotations. The annotation profiles of apaQTLs show more similarity to that of unexplained eQTLs than to that of explained eQTLs. Error bars represent 95% confidence intervals from bootstraps (Online Methods).

is particularly noteworthy because nearly all genetic effects on ribosome occupancy have been proposed to be mediated by effects on mRNA expression (Battle et al., 2015). 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 (Fig 4B, Supplementary Fig. 15), but is not associated with *EIF2A* mRNA levels (Fig. 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 (Mayr, 2017), 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 (Yamashita & Takeuchi, 2017), or differential rates of translation re-initiation at the end of a translation cycle (Rogers, Böttcher, Traulsen, & Greig, 2017).

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. In

both studies, the authors proposed that some genetic effects on protein expression levels were mediated by changes in the protein sequence, which would manifest post-translationally. Our finding reveals yet another complex mode of genetic regulation of protein expression level by APA (e.g. perhaps by affecting recruitment of interacting proteins). Thus, these findings provide clear evidence that APA can affect protein expression levels without affecting gene expression levels through multiple regulatory pathways.

APA mediates genetic effects on complex traits

Lastly, we hypothesized that genetic variation may impact disease risk through APA. Indeed, 19.3% of apaQTLs (including SNPs in LD) are significantly associated with at least one trait in the UCSC GWAS catalog (Kent et al., 2002). Interestingly, an apaQTL in the C10orf88 gene (rs7904973) has been associated with increased LDL cholesterol (Klarin et al., 2018), suggesting that eQTLs mediated by APA can impact organismal phenotypes. APA is complex regulatory mechanism relevant to our understanding of how genetic variation can affect disease; therefore, comprehensive maps of apaQTLs can enhance our ability to interpret GWAS loci, particularly when the implicated variants are not eQTLs (Joehanes et al., 2017; Lee et al., 2018). 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. (Swaminathan et al., 2015) Interestingly, multiple myeloma is among the cancer types in which widespread dysregulation of intronic APA has been documented previously (Singh et al., 2018; Lee et al., 2018).

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 two important shortcomings. Firstly, the study of inter-individual variation in PAS usage have been mostly restricted to APA in the 3′ UTRs (L. Li et al., 2019; Yoon et al., 2012; Yang et al., 2019), 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.

To overcome these shortcomings, we applied 3′ Seq to total and nuclear cell fractions separately to directly measure PAS usage including that of PAS in intronic regions. These data allowed us to study the

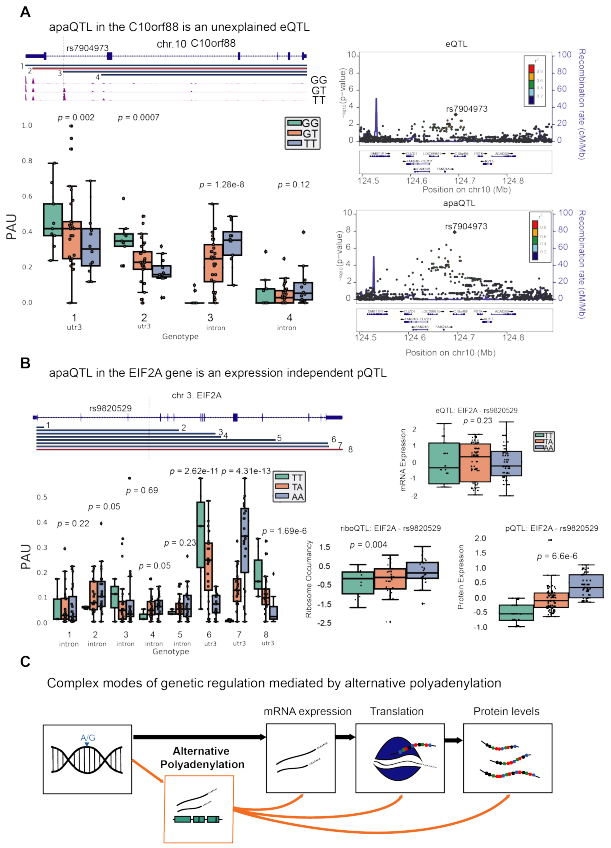

Figure 4: **(A)** Example of an apaQTL that is also an unexplained eQTL in *C10orf88*. *(Top Left)* Gene track and identified PAS in the *C10orf88* gene. The red bar corresponds to the isoform most strongly associated with the apaQTL. The vertical dotted line represents the position of the strongest apaQTL SNP. *(Bottom Left)* Polyadenylation site usage (PAU) at each PAS by genotype listed according to isoform order above. (*Right)* Locus zoom plot for eQTL and apaQTL associations. Interestingly, the top associated SNP, rs7904973, has been linked to increased LDL cholesterol through GWAS (Klarin et al., 2018). **(B)***(Top Left)* Gene track and identified PAS in the *EIF2A* gene. *(Bottom Left)* Polyadenylation site usage (PAU) at each PAS by genotype listed according to isoform order above. *(Right)* Boxplots showing normalized mRNA expression, ribosome occupancy, and protein expression for *EIF2A* by genotype at the apaQTL SNP (rs9820529). **(C)** Diagram representing multiple pathways by which genetic variation mediates gene regulation through alternative polyadenylation.

effects of genetic variation on polyadenylation at multiple stages of the mRNA life cycle, 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, or post-transcriptionally by affecting binding of miRNAs or RNA-binding proteins and consequently isoform decay. However, we found that the vast majority of genetic variants that affect PAS usage ratio in total mRNA, were also found to have highly 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 very 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 handful 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 Fig. 9) (L. Li et al., 2019). 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 (Oh et al., 2017), and RNAP II pausing (Fusby et al., 2016). Indeed, recent studies have reported that sequence and chromatin context can pause or slow down RNAP II elongation across the gene body (Mayer et al., 2015), suggesting that variation in RNAP II pausing may impact PAS choice (Fusby et al., 2016). For example, in Drosophila melanogaster, paused RNAP II promotes the recruitment of ELAV on the pre-mRNA, which prevents usage of a proximal PAS (Oktaba et al., 2015). Interestingly, the human ortholog, ELAVL1 has been implicated in mRNA localization and may influence APA through competition for binding with other factors (Neve, Patel, Wang, Louey, & Furger, 2017; Berkovits & Mayr, 2015; Dai, Zhang, & Makeyev, 2012).

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 (Tian & Manley, 2017; Vasudevan, Peltz, & Wilusz, 2002). 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. Overall, 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 findings reveal several complex modes of genetic regulation for both gene expression and protein expression levels by APA (Fig. 4C). 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.

Methods

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 (International HapMap Consortium, 2005; Moll, Ante, Seitz, & Reda, 2014). Details for each cell line are found in Supplementary Table 2. 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.

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 (Mayer & Churchman, 2016). 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 Supplementary Table 2. We checked the quality of the collected RNA using a nanodrop. RNA concentrations and absorbance levels from the collection are in Supplementary Table 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 (Mayer & Churchman, 2016). 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 Fig. 16).

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 (Moll et al., 2014) 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.

3′ Sequencing data processing

We mapped 3′ Seq reads to hg19 (Church et al., 2011) using STAR RNA-seq aligner (Dobin et al., 2013) using default settings with the WASP mode to filter out reads mapping with allelic bias (van de Geijn, McVicker, Gilad, & Pritchard, 2015). 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 (Tian, Hu, Zhang, & Lutz, 2005; Sheppard, Lawson, & Zhu, 2013; Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000). We verified the individual identity of all bam files using VerifyBamID (Jun et al., 2012). 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 Supplementary Table 1 (Supplementary Fig. 17).

Identification 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 (Kent et al., 2002). When a region mapped to multiple genes we used a hierarchical model, similar to the method used by Lin et al. (Lin et al., 2012) 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 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 (Wang et al., 2018)(Supplementary Fig. 5).

PAS Signal site enrichment and locations

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 (Beaudoing et al., 2000). 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 (Beaudoing et al., 2000).

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 (Liao, Smyth, & Shi, 2014). We ran the leafcutter_ds.R script on chromosomes 1-22 separately using the cellular fraction label as the sample group identifier (Y. I. Li et al., 2018). 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.

Relationship with nascent transcription

We used 30 min 4su data and RNA decay measurements collected in the same panel of LCLs as used in this study. RNA decay data was originally collected and processed in Pai et al. 2012. (Pai et al., 2012) The 4su

data collection and processing can be found in Li et al. 2016(Y. I. Li et al., 2016). We used RNA sequencing data collected in the same LCLs as used in this study. The data collection information can be found in Pickrell et al, 2010 and further processing can be found in Li et al. 2016 (Pickrell et al., 2010; Y. I. Li et al., 2016). We computed a nascent transcription phenotype for each gene as the normalized 4su expression level over the sum of the RNA expression and 4su expression level. We calculated the correlation between this value as well as nuclear $3^{'}$Seq read counts for each gene divided by the sum of total $3^{'}$ Seq read counts and nuclear $3^{'}$ Seq read counts for each gene. We also calculated the correlation between the difference in the number of PAS identified both fractions and the nascent transcription phenotype using the summary lm function in R.

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 (Supplementary Fig. 8). We included the top 4 PCs as well as the library preparation batch as the covariates. The top four PCs correlate most strongly with the cell count at collection (Supplementary Fig. 8). We used the same genotypes from Li et al. 2016(Y. I. Li et al., 2016), available at http://eqtl.uchicago.edu/jointLCL/genotypesYRI.gen.txt.gz (Y. I. Li et al., 2016). 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 (Ongen, Buil, Brown, Dermitzakis, & Delaneau, 2016). 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.

Association of apaQTLs with chromatin states

We downloaded the GM12878 chromatin HMM annotations for Hg19 from the UCSC table browser (Kent et al., 2002). We overlapped the eQTLs identified and published in Li et al. 2016(Y. I. Li et al., 2016) 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 (Fig. 3d and Supplementary Fig. 18).

apaQTL overlap with eQTLs

We obtained the set of explained and unexplained eQTLs from Li et al. 2016 (Y. I. Li et al., 2016). 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(Y. I. Li et al., 2016) 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 Fig. 19).

apaQTLs overlap with protein specific QTLs

The list of protein specific QTL genes can be found in the supplementary information from Battle et al. 2015(Battle et al., 2015). 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 or eGene. We tested for significant enrichment with one sided Wilcoxon rank sum tests.

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 (Y. I. Li et al., 2016). We used normalized RNA expression, RiboSeq values, and protein levels, published in Li et al. 2016 (Y. I. Li et al., 2016).

apaQTL overlap with GWAS Catalog

We downloaded the CRCh37hg19 GWAS catalog for UCSC table browser (Kent et al., 2002). We identified SNPs in LD with the nuclear apaQTLs using the LDproxy tool from LDlink with YRI as the population (Machiela & Chanock, 2015). We filtered all results to SNPs with an r2 greater than 0.9. We overlapped the full set with the GWAS catalog using pybedtools.

Acknowledgements

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.

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. 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.

Competing interests

The authors declare no competing interest.

Data and material Availability

Fastq files and PAS annotations are available at GEO under accession GSE138197. All reproducible scripts can be found at `https://brimittleman.github.io/apaQTL/`

References

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., & Gilad, Y. (2015, February). Genomic variation. Impact of regulatory variation from RNA to protein. *Science (New York, N.Y.)*, *347*(6222), 664-667. doi: 10.1126/science.1260793

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J.-M., & Gautheret, D. (2000, July). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research*, *10*(7), 1001-1010.

Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., ... Dreyfuss, G. (2012, July). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. *Cell*, *150*(1), 53-64. doi: 10.1016/j.cell.2012.05.029

Berkovits, B. D., & Mayr, C. (2015, June). Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature*, *522*(7556), 363-367. doi: 10.1038/nature14321

Bonder, M. J., Smail, C., Gloudemans, M. J., Frésard, L., Jakubosky, D., D'Antonio, M., ... Stegle, O. (2019, October). Systematic assessment of regulatory effects of human disease variants in pluripotent cells. *bioRxiv*, 784967. doi: 10.1101/784967

Chick, J. M., Munger, S. C., Simecek, P., Huttlin, E. L., Choi, K., Gatti, D. M., ... Gygi, S. P. (2016, June). Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, *534*(7608), 500-505. doi: 10.1038/nature18270

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., ... Cotsapas, C. (2017, April). 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. doi: 10.1038/ng.3795

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., ... Hubbard, T. (2011, July). Modernizing reference genome assemblies. *PLoS biology*, *9*(7), e1001091. doi: 10.1371/journal.pbio.1001091

Dai, W., Zhang, G., & Makeyev, E. V. (2012, January). RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Research*, *40*(2), 787-800. doi: 10.1093/nar/gkr783

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., ... Pritchard, J. K. (2012, February). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, *482*(7385), 390-394. doi: 10.1038/nature10808

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013, January). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21. doi: 10.1093/bioinformatics/bts635

Fusby, B., Kim, S., Erickson, B., Kim, H., Peterson, M. L., & Bentley, D. L. (2016, January). Coordination of RNA Polymerase II Pausing and 3′ End Processing Factor Recruitment with Alternative Polyadenylation. *Molecular and Cellular Biology*, *36*(2), 295-303. doi: 10.1128/MCB.00898-15

Ha, K. C. H., Blencowe, B. J., & Morris, Q. (2018, March). QAPA: A new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology*, *19*(1), 45. doi: 10.1186/s13059-018-1414-4

International HapMap Consortium. (2005, October). A haplotype map of the human genome. *Nature*, *437*(7063), 1299-1320. doi: 10.1038/nature04226

Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.-x., Nguyen, Q. T., … Munson, P. J. (2017, January). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology*, *18*(1), 16. doi: 10.1186/s13059-016-1142-6

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., … Kang, H. M. (2012, November). 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. doi: 10.1016/j.ajhg.2012.09.004

Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L., & Dreyfuss, G. (2010, December). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, *468*(7324), 664-668. doi: 10.1038/nature09479

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002, January). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996-1006. doi: 10.1101/gr.229102

Klarin, D., Damrauer, S. M., Cho, K., Sun, Y. V., Teslovich, T. M., Honerlaw, J., … Assimes, T. L. (2018, November). Genetics of blood lipids among ˜300,000 multi-ethnic participants of the Million Veteran Program. *Nature Genetics*, *50*(11), 1514-1523. doi: 10.1038/s41588-018-0222-9

Lee, S.-H., Singh, I., Tisdale, S., Abdel-Wahab, O., Leslie, C. S., & Mayr, C. (2018, September). Widespread intronic polyadenylation inactivates tumor suppressor genes in leukemia. *Nature*, *561*(7721), 127-131. doi: 10.1038/s41586-018-0465-8

Li, L., Gao, Y., Peng, F., Wagner, E. J., & Li, W. (2019, March). Genetic Basis of Alternative Polyadenylation

is an Emerging Molecular Phenotype for Human Traits and Diseases. *bioRxiv*, 570176. doi: 10.1101/570176

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018, January). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, *50*(1), 151-158. doi: 10.1038/s41588-017-0004-9

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., … Pritchard, J. K. (2016, April). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*(6285), 600-604. doi: 10.1126/science.aad9417

Liao, Y., Smyth, G. K., & Shi, W. (2014, April). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923-930. doi: 10.1093/bioinformatics/btt656

Lin, Y., Li, Z., Ozsolak, F., Kim, S. W., Arango-Argoty, G., Liu, T. T., … John, B. (2012, September). An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Research*, *40*(17), 8460-8471. doi: 10.1093/nar/gks637

Machiela, M. J., & Chanock, S. J. (2015, November). 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. doi: 10.1093/bioinformatics/btv402

Mayer, A., & Churchman, L. S. (2016, April). Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing. *Nature Protocols*, *11*(4), 813-833. doi: 10.1038/nprot.2016.047

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., … Churchman, L. S. (2015, April). Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell*, *161*(3), 541-554. doi: 10.1016/j.cell.2015.03.010

Mayr, C. (2016, March). Evolution and Biological Roles of Alternative 3'UTRs. *Trends in Cell Biology*, *26*(3), 227-237. doi: 10.1016/j.tcb.2015.10.012

Mayr, C. (2017, November). Regulation by 3′-Untranslated Regions. *Annual Review of Genetics*, *51*(1), 171-194. doi: 10.1146/annurev-genet-120116-024704

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., … Pritchard, J. K. (2013, November). Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*, *342*(6159), 747-749. doi: 10.1126/science.1242429

Moll, P., Ante, M., Seitz, A., & Reda, T. (2014, November). QuantSeq 3′ mRNA sequencing for RNA

quantification. *Nature Methods*, *11*, 972. doi: 10.1038/nmeth.f.376

Neve, J., Patel, R., Wang, Z., Louey, A., & Furger, A. M. (2017, April). Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biology*, *14*(7), 865-890. doi: 10.1080/15476286.2017.1306171

Oh, J.-M., Di, C., Venters, C. C., Guo, J., Arai, C., So, B. R., ... Dreyfuss, G. (2017, November). U1 snRNP telescripting regulates a size–function-stratified human genome. *Nature Structural & Molecular Biology*, *24*(11), 993-999. doi: 10.1038/nsmb.3473

Oktaba, K., Zhang, W., Lotz, T. S., Jun, D. J., Lemke, S. B., Ng, S. P., ... Hilgers, V. (2015, January). ELAV Links Paused Pol II to Alternative Polyadenylation in the Drosophila Nervous System. *Molecular Cell*, *57*(2), 341-348. doi: 10.1016/j.molcel.2014.11.024

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016, May). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, *32*(10), 1479-1485. doi: 10.1093/bioinformatics/btv722

Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.-B., ... Gilad, Y. (2012, October). The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels. *PLoS Genetics*, *8*(10), e1003000. doi: 10.1371/journal.pgen.1003000

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... Pritchard, J. K. (2010, April). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768-772. doi: 10.1038/nature08872

Rogers, D. W., Böttcher, M. A., Traulsen, A., & Greig, D. (2017, June). Ribosome reinitiation can explain length-dependent translation of messenger RNA. *PLOS Computational Biology*, *13*(6), e1005592. doi: 10.1371/journal.pcbi.1005592

Sheppard, S., Lawson, N. D., & Zhu, L. J. (2013, October). Accurate identification of polyadenylation sites from 3′ end deep sequencing using a naïve Bayes classifier. *Bioinformatics*, *29*(20), 2564-2571. doi: 10.1093/bioinformatics/btt446

Shi, Y. (2012, December). Alternative polyadenylation: New insights from global analyses. *RNA (New York, N.Y.)*, *18*(12), 2105-2117. doi: 10.1261/rna.035899.112

Singh, I., Lee, S.-H., Sperling, A. S., Samur, M. K., Tai, Y.-T., Fulciniti, M., ... Leslie, C. S. (2018, December). Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nature Communications*, *9*(1), 1716. doi: 10.1038/s41467-018-04112-z

Swaminathan, B., Thorleifsson, G., Jöud, M., Ali, M., Johnsson, E., Ajore, R., ... Nilsson, B. (2015, May).

Variants in *ELL2* influencing immunoglobulin levels associate with multiple myeloma. *Nature Communications*, *6*, 7213. doi: 10.1038/ncomms8213

Tian, B., Hu, J., Zhang, H., & Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, *33*(1), 201-212. doi: 10.1093/nar/gki158

Tian, B., & Manley, J. L. (2017, January). Alternative polyadenylation of mRNA precursors. *Nature Reviews. Molecular Cell Biology*, *18*(1), 18-30. doi: 10.1038/nrm.2016.116

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015, November). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, *12*(11), 1061-1063. doi: 10 .1038/nmeth.3582

Vasudevan, S., Peltz, S. W., & Wilusz, C. J. (2002). Non-stop decay—a new mRNA surveillance pathway. *BioEssays*, *24*(9), 785-788. doi: 10.1002/bies.10153

Wang, R., Nambiar, R., Zheng, D., & Tian, B. (2018, January). PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Research*, *46*(D1), D315-D319. doi: 10.1093/nar/gkx1000

Yamashita, A., & Takeuchi, O. (2017, April). Translational control of mRNAs by 3′-Untranslated region binding proteins. *BMB reports*, *50*(4), 194-200. doi: 10.5483/bmbrep.2017.50.4.040

Yang, Y., Zhang, Q., Miao, Y.-R., Yang, J., Yang, W., Yu, F., … Gong, J. (2019, September). SNP2APA: A database for evaluating effects of genetic variants on alternative polyadenylation in human cancers. *Nucleic Acids Research*. doi: 10.1093/nar/gkz793

Yoon, O. K., Hsu, T. Y., Im, J. H., & Brem, R. B. (2012, August). Genetics and Regulatory Impact of Alternative Polyadenylation in Human B-Lymphoblastoid Cells. *PLOS Genetics*, *8*(8), e1002882. doi: 10.1371/journal.pgen.1002882