Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees

Briana E Mittleman[1], Sebastian Pott[2], Shane Warland[3], Kenneth Barr[3], Claudia Cuevas[3], Yoav Gilad[2,3]*

[1]Genetics, Genomics and Systems Biology, University of Chicago, Chicago, United States; [2]Department of Human Genetics, University of Chicago, Chicago, United States; [3]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, United States

Abstract While comparative functional genomic studies have shown that inter-species differences in gene expression can be explained by corresponding inter-species differences in genetic and epigenetic regulatory mechanisms, co-transcriptional mechanisms, such as alternative polyadenylation (APA), have received little attention. We characterized APA in lymphoblastoid cell lines from six humans and six chimpanzees by identifying and estimating the usage for 44,432 polyadenylation sites (PAS) in 9518 genes. Although APA is largely conserved, 1705 genes showed significantly different PAS usage (FDR 0.05) between species. Genes with divergent APA also tend to be differentially expressed, are enriched among genes showing differences in protein translation, and can explain a subset of observed inter-species protein expression differences that do not differ at the transcript level. Finally, we found that genes with a dominant PAS, which is used more often than other PAS, are particularly enriched for differentially expressed genes.

***For correspondence:** gilad@uchicago.edu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 20

Received: 28 August 2020
Accepted: 12 February 2021
Published: 17 February 2021

Reviewing editor: Graham Coop, University of California, Davis, United States

Introduction

Humans and our close primate relatives exhibit a striking array of phenotypic diversity despite sharing homologous proteins with nearly identical amino acid sequences (***King and Wilson, 1975***). Understanding how this diversity is propagated from genomic sequence to mRNA and then to protein necessitates an understanding of the regulatory mechanisms that occur before, during, and after transcription. Studying gene regulatory features in humans and other primates has long provided opportunities to understand genome evolution and function. For example, studies comparing patterns of epigenetic marks in primates have provided mechanistic explanations that link genetic variation and divergence to differences in gene expression levels (***Banovich et al., 2014***; ***Cain et al., 2011***; ***McVicker et al., 2013***; ***Pai et al., 2011***). Although many studies have focused on inter-species differences in the regulation of gene expression, fewer studies have addressed isoform-level variation, which contributes to differences in mRNA, translation, and protein levels between species (***Blekhman et al., 2010***; ***Calarco et al., 2007***; ***Pai et al., 2016***).

The main mechanisms that contribute to mRNA isoform diversity are alternative splicing and alternative polyadenylation (APA). Alternative splicing produces different combinations of coding sequences in mature mRNA and protein. APA occurs at genes that have more than one polyadenylation site (PAS) and can result in mRNAs with different coding sequences or variable 3′ UTR lengths. Like alternative splicing, APA that occurs within the gene body can affect protein sequence and function. (***Lee et al., 2018***; ***Pan et al., 2006***; ***Sandberg et al., 2008***; ***Tian and Manley, 2017***;

Vasudevan et al., 2002; *Yao et al., 2018*). Unlike with alternative splicing, a PAS in a coding region will lead to a truncated isoform rather than a different combination of included exons. APA that occurs outside of the coding sequence, in the 3′ UTRs, can lead to differential inclusion of protein-binding motifs that can affect translational efficiency, mRNA stability, and mRNA localization (*Mayr, 2017*; *Tian and Manley, 2017*). Yet, despite its potential to produce tremendous variation in mRNA and protein regulation, few studies have explored the contribution of APA to regulatory divergence between species. Indeed, our current understanding of APA conservation in mammals comes from few comparative studies of humans and rodents (*Ara et al., 2006*; *Wang et al., 2018a*). However, these studies used sequence conservation rather than direct measurements of PAS usage to characterize APA (*Wang et al., 2018a*). Thus, it remains possible that many mammalian PAS are functionally divergent despite having similar sequences.

To gain insight into APA conservation in humans and chimpanzees and understand how differences in APA contribute to gene regulation, we performed 3′ sequencing (3′ Seq) of mRNA isolated from nuclei collected from human and chimpanzee lymphoblastoid cell lines (LCLs). We integrated PAS usage measurements with RNA-sequencing (RNA-seq) data collected from the same cell lines to understand the relationship between APA and gene expression levels. Finally, we used ribosome profiling and protein measurements previously collected in the same panel of human and chimpanzee LCLs to explore the effects of APA on protein levels (*Khan et al., 2013*; *Wang et al., 2018b*). We took this approach because an understanding of how APA isoform usage varies among primates could help explain why some human and chimpanzee genes are differentially expressed at either the mRNA or protein levels, but not both.

Results

Describing APA in human and chimpanzee LCLs

We performed 3′ Seq of mRNA from six human and six chimpanzee LCLs, which we have previously used to study a variety of other functional genomic phenotypes, such as ribosome profiling to infer translation levels and mass spectrometry to measure protein levels (*Cain et al., 2011*; *Khan et al., 2013*; *Wang et al., 2018b*; *Zhou et al., 2014*). We collected mRNA separately from whole cells and isolated nuclei. The two cellular fractions serve as biological replicates, which we mainly used to examine the quality of our data (see 'Materials and methods'). By collecting data from isolated nuclei, we were able to capture polyadenylated transcripts before they became undetectable due to other regulatory processes, such as isoform-specific decay (*Mittleman et al., 2020*).

We mapped human 3′ Seq reads to the GRCh38 reference genome (*Schneider et al., 2017*) and chimpanzee 3′ Seq reads to the panTro6 reference genome (*Chimpanzee Sequencing and Analysis Consortium, 2005*) (see 'Materials and methods'). 3′ Seq relies on a poly(dT) primer to target the poly(A) tail of mRNA molecules; however, it can also misprime by binding a sequence of genomic adenines. To account for mispriming of off-target genomic sequences, we removed reads that mapped to genomic regions containing $\geq$70% adenine or six consecutive adenine bases in the 10 bp directly upstream of the mapped location (*Mittleman et al., 2020*; *Sheppard et al., 2013*; *Tian et al., 2005*) (see 'Materials and methods'). In addition, we treated all ambiguous nucleotide positions as adenines to ensure that differences in reference genome quality did not bias the detection of PAS or mispriming events (see 'Materials and methods'). As expected, the filtered aligned sequences, in both species, were enriched at transcription end sites and showed a similar distribution along orthologous 3′ UTRs (see 'Materials and methods', *Figure 1—figure supplement 1*). Next, we used a custom peak calling method to ascertain PAS in humans and chimpanzees separately (see 'Materials and methods').

To compare PAS usage across species, we needed to identify the orthologous genomic regions of all PAS in our dataset, regardless of the species in which they were originally annotated. As we were unable to confidently identify orthologous PAS at base pair resolution (inferring synteny at base pair resolution in non-coding regions is challenging; Broad Institute *Lindblad-Toh et al., 2011*), we extended each PAS by 100 bp upstream and downstream. We then used a reciprocal liftover pipeline to obtain an inclusive set of PAS regions with which we could confidently compare PAS usage between species (see 'Materials and methods'). Prior to the filtering described below, we identified 445,944 orthologous regions.

To quantify PAS usage, we first assigned each PAS to a gene using the hg38 RefSeq annotation (**Pruitt et al., 2004**). We then computed the usage for each PAS in each individual as the fraction of reads mapping to one PAS over the total number of reads mapping to any PAS for the same gene (**Figure 1—figure supplement 2**). We excluded PAS in lowly expressed genes ($\log_2$(CPM) ≤ 2 in four or more individuals) or with less than 5% usage, as measurements from sparse data are highly susceptible to random error (see 'Materials and methods'). We observed a strong correlation between PAS usage in mRNA from the nuclear and total cell fractions in all but one cell line (human NA18499; **Figure 1—figure supplement 3**), which we subsequently excluded from the study. We re-identified PAS after removing all data from NA18499 and re-quantified PAS usage using nuclear 3′ Seq data from five human and six chimpanzee LCLs. Using this analysis pipeline, we identified a total of 44,432 PAS in 9518 genes, which we used for all downstream analyses. On a genome-wide scale, we found that mean PAS usage is highly correlated between species (Pearson's correlation, 0.9, $p<2.2\times10^{-16}$, **Figure 1—figure supplement 4**). However, as expected, 41.8% of the variation in PAS usage (as explained by the top principal component of the data) is highly correlated with species (Pearson's correlation 0.99, $p=2.95\times10^{-8}$, **Figure 1—figure supplement 5**), indicating substantial divergence in PAS usage.

We used a number of analyses to confirm that our ability to detect PAS was not biased by gene expression level or species. If our ability to detect PAS was biased by gene expression, we might expect a positive correlation between gene expression level and the number of PAS we detected. In our data, the number of PAS per gene is negatively correlated with gene expression level in both species (**Figure 1—figure supplement 6**; human: Pearson's correlation -0.17, $p<2.2\times10^{-16}$; chimpanzee: Pearson's correlation -0.19, $p<2.2\times10^{-16}$). If our ability to detect PAS were biased by species, we would expect to identify more PAS per gene in one species over the other. This is neither the case genome-wide nor when we test each gene independently. We identified, on average, 3.87 PAS per gene in humans and 3.46 PAS per gene in chimpanzees. On average, per gene, the number of PAS in human minus the number of PAS in chimpanzee is 0.39 with a median value of 0 (**Figure 1—figure supplement 7**). Moreover, as expected, the physical distribution of PAS across genes is conserved, with the majority of PAS located in 3′ UTRs (17,688, 40% in chimpanzee; and 17,620, 40% in human) and a considerable proportion located in introns (14,095, 32% in chimpanzee; and 14,119, 32% in human) (**Figure 1A**).

To assess sequence conservation in PAS regions, we downloaded PhyloP scores computed over 100 vertebrate genomes from the UCSC Genome Browser and calculated mean PhyloP scores in PAS regions. Higher mean PhyloP scores correspond to regions of higher sequence conservation and thus slower evolution (**Pollard et al., 2010**). Overall, sequence elements at PAS are more conserved than 200 base pair surrounding regions 1 kb on either side of the PAS (**Figure 1B**, Wilcoxon rank sum test, $p<2.2\times10^{-16}$). This pattern also holds independently for PAS in all genic locations other than introns (**Figure 1—figure supplement 8**). By repeating these analyses with PhyloP score computed over 20 vertebrates, we show the PAS are also conserved compared to the surrounding regions in the mammalian lineage (**Figure 1—figure supplement 9**).

We identified 302 and 357 human- and chimpanzee-specific PAS, respectively (see 'Materials and methods'). Suggesting biological significance, compared to the genes in which we identified a PAS in both species, the genes with species-specific PAS are enriched for a number of general cellular processes (**Supplementary file 1**, see 'Materials and methods'). It has been previously shown that most PAS are directly preceded by 1 of 12 annotated sequence motifs that recruit cleavage and polyadenylation machinery to mRNA molecules as they are transcribed (**Beaudoing et al., 2000**). We asked if creation or disruption of a signal site motif could be responsible for species-specific PAS by mapping signal site motifs in both human and chimpanzee for each PAS region. Although human and chimpanzee PAS regions are equally likely to contain each of the 12 annotated signal sites (**Figure 1C**), only the top two most commonly used motifs, AATAAA and ATTAAA, are associated with increased PAS usage (**Figure 1—figure supplement 10**). Not only have other studies revealed a similar preference for these two motifs, cryo-electric microscopy analysis of the recognition machinery suggests that the mutation from the canonical signal site to the second most used site requires a smaller RNA rearrangement than other variations of the motif (**Beaudoing et al., 2000**; **Sun et al., 2018**). Thus, we considered only the presence or absence of these two motifs in subsequent analyses. We classified sites as having a species-specific signal site if we only identified the AATAAA or ATTAAA motif in one species. Of the 302 human-specific PAS, 14 have human-specific

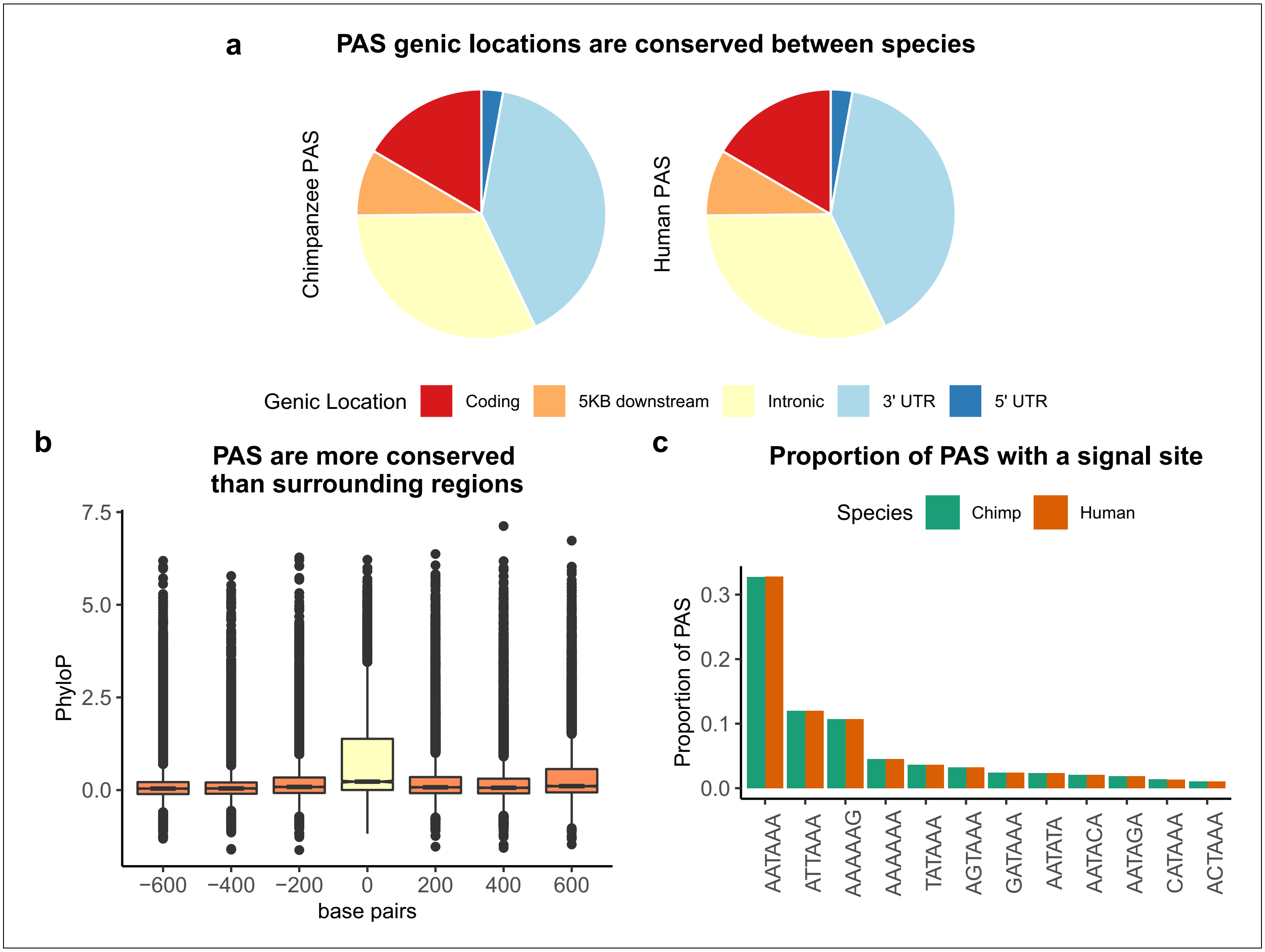

Figure 1. Sequence conservation of polyadenylation sites (PAS) between humans and chimpanzees. (**a**) Genic locations for 44,074 PAS identified in chimpanzee (left) and 44,130 PAS identified in human (right). (**b**) Mean PhyloP scores for PAS regions (yellow) as well as three 200 bp regions upstream and downstream (orange). (**c**) Proportion of human and chimpanzee PAS regions with each of the 12 annotated signal site motifs from ***Beaudoing et al., 2000***.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Density of merged human and chimpanzee 3′ sequencing (3′ Seq).
Figure supplement 2. Model representation of usage calculation.
Figure supplement 3. NA18499 removed from analysis due to low correlation between fractions.
Figure supplement 4. Polyadenylation sites (PAS) usage is highly correlated across species.
Figure supplement 5. Variation in polyadenylation sites (PAS) usage.
Figure supplement 6. Polyadenylation sites (PAS) detection likely not biased by expression level.
Figure supplement 7. Polyadenylation sites (PAS) detection likely not biased by species.
Figure supplement 8. ***Figure 1B*** separated by genic location.
Figure supplement 9. Polyadenylation sites (PAS) regions are conserved in the mammalian lineage.
Figure supplement 10. PAS with AATAAA and ATTAAA are used more often.
Figure supplement 11. Chimp-specific polyadenylation sites (PAS) likely due to loss of signal site in human lineage.
Figure supplement 12. Reciprocal liftover pipeline.

signal sites and 6 have chimpanzee-specific signal sites. Of the 357 chimpanzee-specific PAS, 24 have a chimpanzee-specific signal site and 6 have a human-specific signal site. These numbers are small; still, species-specific signal sites are more abundant than expected by chance among species-specific PAS in human (5.7×, hypergeometric test, $p=2.30\times10^{-7}$) and in chimpanzee (8.3×, hypergeometric test, $p=3.2\times10^{-15}$), suggesting that signal site changes can explain a subset of differences in PAS usage. For example, we identified a chimpanzee-specific PAS about 1 kb upstream of a PAS used in both species in the 3′ UTR of *MAN2B2*. The ancestral signal site conserved in chimpanzee is AATAAA; however, there has been a T to C transition in the human lineage (**Blanchette et al., 2004**; **Figure 1—figure supplement 11**). This transition is likely responsible for the loss of PAS in humans. The MAN2B2 gene involved with lysosomal degradation of glycoproteins, and in 2019 a physician diagnosed a patient with immune deficiency as a result of a loss of function mutation in the gene (**Verheijen et al., 2020**).

Characterizing inter-species differences in PAS usage

While a few hundred PAS are species-specific, the majority of PAS (98.5%) were identified in both species. We thus sought to characterize quantitative differences in APA patterns between human and chimpanzee by estimating the difference in the usage of individual PAS in each species. To do so, we used the leafcutter differential splicing tool (**Li et al., 2018**), which allowed us to test for differences in normalized PAS usage fractions while accounting for gene structure (see ‘Materials and methods’). Using this approach, at a false discovery rate (FDR) of 5% we identified 2342 PAS (in 1705 genes) whose usage differs by 20% or more between the species (**Figure 2A**). We applied an arbitrary effect size cutoff to focus on larger inter-species differences, which are more likely to have functional consequences. The list of all PAS whose usage differs between the species, regardless of the effect size, is available in **Supplementary file 2**.

To better understand the mechanisms that underlie inter-species differences in PAS usage, and the potential functional impact of such differences, we considered the APA data in different contexts. First, we noticed that the spatial distribution of differentially used PAS reflects the distribution of all PAS; namely, differentially used PAS are most often located in 3′ UTRs, followed by introns (**Figure 2—figure supplement 1**). Within the 3′ UTR, however, differentially used PAS are more frequently the first ones compared with PAS that are used similarly in the two species (**Figure 2—figure supplement 2**, difference in proportion test, p=0.0015). This pattern is intriguing, because changes in the usage of the first PAS in the 3′ UTR may have the largest overall impact on the transcript length, and hence potentially the largest functional impact. However, it is also possible that we are more likely to detect differences in the usage in the first PAS in the 3′ UTR because this site is transcribed earlier, and our estimate of usage is relative to all other sites in each gene.

We therefore sought evidence that differences in PAS usage may have functional consequences. In a previous study, we identified genetic variants associated with variation in PAS usage (apaQTLs) in a panel of 52 human LCLs (**Mittleman et al., 2020**). We found that genes with inter-species differences in PAS usage are highly enriched for apaQTLs (160, empirical p-value based on 10,000 permutations = 0.001, **Figure 2—figure supplements 3** and **1**, 3×; hypergeometric test, p=0.0009). This observation indicates that inter-species differences in APA usage can often be found in genes whose regulation varies also at the population level, generally suggesting relaxation of evolutionary constraint on the regulation of such genes. We next considered sequence divergence at PAS by obtaining PhyloP scores for all PAS flanking regions (200 bp, as explained above). If many changes in PAS usage are genetically controlled, we would expect genomic regions of differentially used PAS to be less conserved than regions containing PAS that have similar usage. Indeed, differentially used sites are enriched for regions with negative mean PhyloP scores (1.02×, hypergeometric test, p=0.02). This observation indicates that sequence divergence is often associated with differences in PAS usage, and that the majority of PAS usage in humans and chimpanzees may be generally conserved due to evolutionary constraint.

We next asked, more specifically, if signal site changes are likely to lead to differences in PAS usage. We addressed this question by performing two analyses. First, we focused on the 82 differentially used PAS with a signal site that is annotated in only one of the species. We found that the presence of a species-specific signal site is associated with increased PAS usage, as might be expected (human enrichment 3.82×, hypergeometric $p=1.37\times10^{-10}$; chimpanzee enrichment 3.02×, $p=3.91\times10^{-8}$). Second, we considered the presence of G/U-rich elements, which are known signals

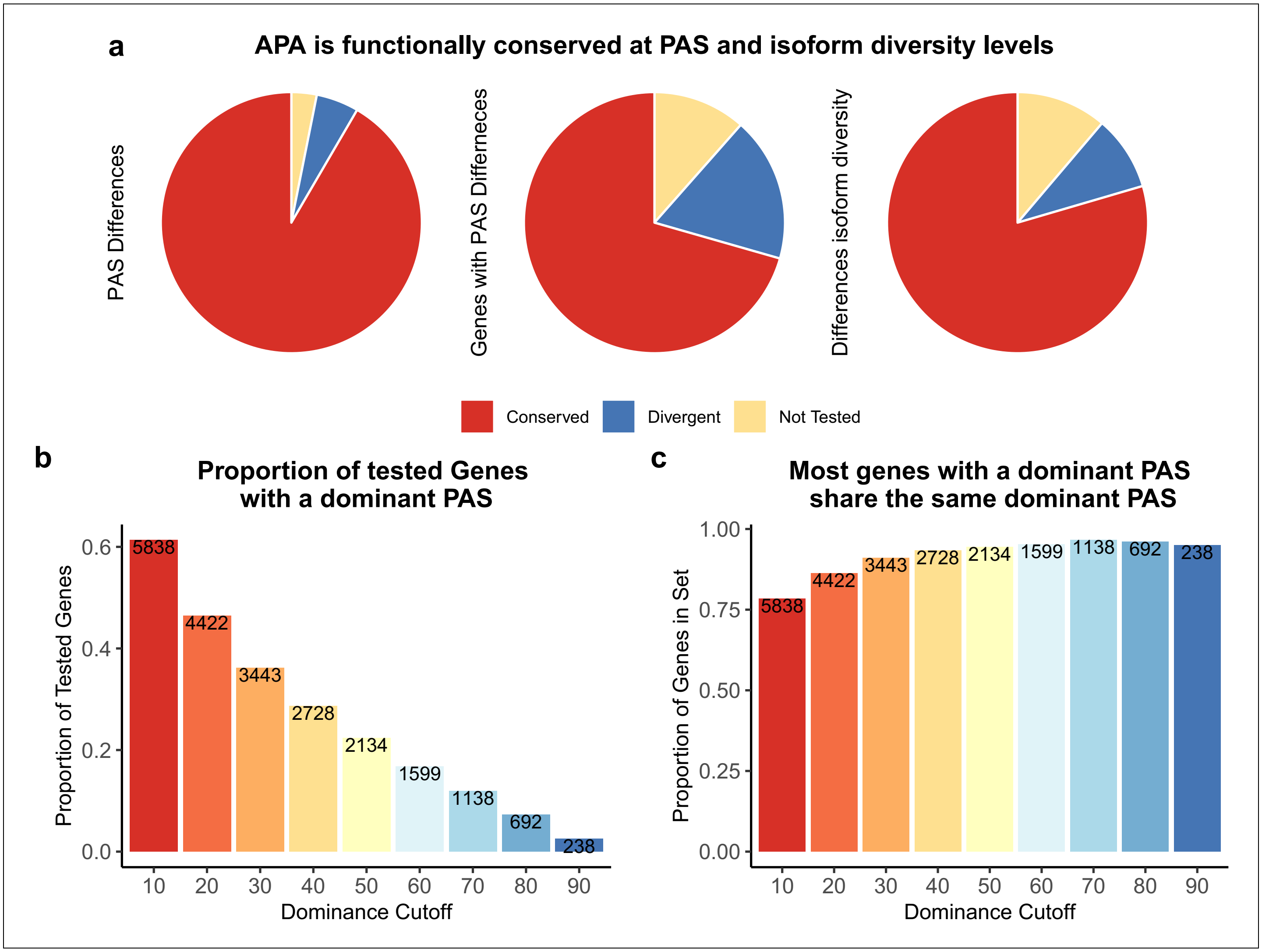

Figure 2. Alternative polyadenylation (APA) is functionally conserved in both species. (**a**) Proportion of polyadenylation sites (PAS) and genes differentially used at PAS and isoform diversity level. (Left) Divergent PAS are the 2342 PAS differentially used at 5% FDR. Conserved are the PAS not differentially used at 5% FDR. Not tested PAS were removed from analysis by leafcutter tool. (Middle) PAS level differentially used PAS reported at the gene level. Divergent genes are the 1705 genes with PAS differentially used at 5% FDR. Conserved genes are the genes with no PAS differentially used at 5% FDR. (Right) Divergent genes are the 881 genes with differences in isoform diversity between species at a 5% FDR. Conserved genes are genes without differences in isoform diversity. Genes with one PAS were not tested. (**b**) Proportion of tested genes with a dominant PAS in either species according to a range of cutoffs. Number of genes are reported in bars. The bars are colored by the dominance cutoff on the X axis. (**c**) Proportion of the number of genes with a dominant in either species that share the top used PAS according to each dominance cutoff. Number of genes with a dominant PAS in either species are reported in bars. The bars are colored by the dominance cutoff on the X axis.
The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Genic location of polyadenylation sites (PAS) differentially used between human and chimpanzee.
Figure supplement 2. Location of polyadenylation sites (PAS) within orthologous 3′ UTRs.
Figure supplement 3. Genes with differentially used polyadenylation sites (PAS) are enriched for genes with apaQTL.
Figure supplement 4. Information content measurement densities.
Figure supplement 5. Relationship between Shannon index and polyadenylation sites (PAS) number.
Figure supplement 6. Relationship between Simpson diversity index and polyadenylation sites (PAS) number.
Figure supplement 7. Intersection between genes with polyadenylation sites (PAS) and isoform diversity differences.
Figure supplement 8. Polyadenylation sites (PAS) that do not lift from human to chimp.

to the molecular machinery for polyadenylation (**Colgan and Manley, 1997**). Specifically, we considered the proportion of uracil bases in the PAS regions. Despite a high correlation in overall uracil content in both species (Pearson's correlation 0.99, $p<2.2\times10^{-16}$), the usage of PAS with greater uracil density in one species is more likely than expected by chance to be upregulated in that species (chimpanzee 1.04× enrichment, p=0.03; human 1.06× enrichment, p=0.03). Though species-specific signal sites explain a modest proportion of inter-species differences in PAS usage, these cases demonstrate the link between sequence evolution and conservation of PAS usage.

The relationship between differences in APA and gene expression

Our analysis to this point indicates that inter-species differences in PAS usage are often genetically controlled, but generally we have not found strong evidence that they are functionally important. We explored this further by considering the APA data in the context of gene expression data that we collected from the same six human and six chimpanzee LCLs (see 'Materials and methods' for data collection procedures and low-level analysis of the RNA-seq data). We found no meaningful correlation between inter-species differences in gene expression levels and changes in polyadenylation site usage (ΔPAU) in 7462 genes for which we had both types of data (Pearson's correlation = -0.06, $p=3.1\times10^{-7}$, **Figure 3A**). We then separately considered the data for the 3' UTR and intronic PAS, because we previously found a different relationship between PAS usage in these genic regions and gene expression levels (**Mittleman et al., 2020**). Indeed, we found that inter-species differences in the usage of intronic and 3' UTR PAS are loosely correlated with differences in expression effect size between the species at an equal magnitude but in opposite directions (**Figure 3B**). Increased usage of intronic sites is correlated with increased expression levels, while increased usage of 3' UTR sites is correlated with decreased expression.

To further investigate the small but significant correlation above, we focused on 3796 genes that were classified as differentially expressed between humans and chimpanzees at 5% FDR (see 'Materials and methods'). We found that genes with at least one differentially used PAS between the species are more likely to be classified as differentially expressed than expected by chance (610 genes, 1.12× enrichment, hypergeometric test, $p=3.18\times10^{-5}$). Within the differentially expressed gene set, the genes with at least one differentially used PAS are enriched for RNA-processing pathways, such as RNA catabolic processes and RNA metabolic processes (**Supplementary file 3**). The genes are also enriched in RNA-processing cellular compartments such as ribosomes and ribonucleoprotein complexes (**Supplementary file 3**, see 'Materials and methods'). Examining the subset of 610 genes, we observed a modest but significant negative correlation between differential expression effect size and ΔPAU when we considered all PAS (Pearson's correlation = -0.15, p=0.0023, **Figure 3C**). Separating the analysis by PAS, genic location revealed, again, an opposite direction of the correlation between gene expression and the usage of either 3' UTR or intronic PAS (**Figure 3D**). These observations are consistent when we use PAS data based on 3' Seq data from whole cells instead of from the nuclear fractions, suggesting that the observed relationship is not due to nuclear export failure (**Figure 3—figure supplement 1**, see 'Materials and methods').

To provide possible mechanistic insight into the relationship between PAS usage and gene expression, we identified AU-rich elements (AREs) in 3' UTRs in both human and chimpanzee. AREs in 3' UTRs have been linked to destabilization of mRNA transcripts and translation repression (**Floor and Doudna, 2016**; **Moore et al., 2014**; **Siegel et al., 2020**). AREs are recognized by a diverse group of binding proteins, leading to multiple models for why the pathway exists (reviewed in **Barreau et al., 2005**). Of note, AREs and the associated binding proteins have been associated with exosome, stress granules, and P-bodies suggesting that AREs are important for response to physiological cell stress signals (**Barreau et al., 2005**; **Kedersha and Anderson, 2002**). We found that the 3' UTRs of genes that show an inter-species difference in 3' UTR PAS usage have a higher number (Wilcoxon test, $p<10^{-16}$, **Figure 3—figure supplement 2**) and density ($p=5.2\times10^{-6}$, **Figure 3—figure supplement 2**) of AREs compared with genes in which the 3' UTR PAS is similarly used in the two species.

Considering overall APA diversity

We explored the relationship between inter-species differences in APA and gene expression by using a different perspective. We hypothesized that we could gain more insight into regulatory

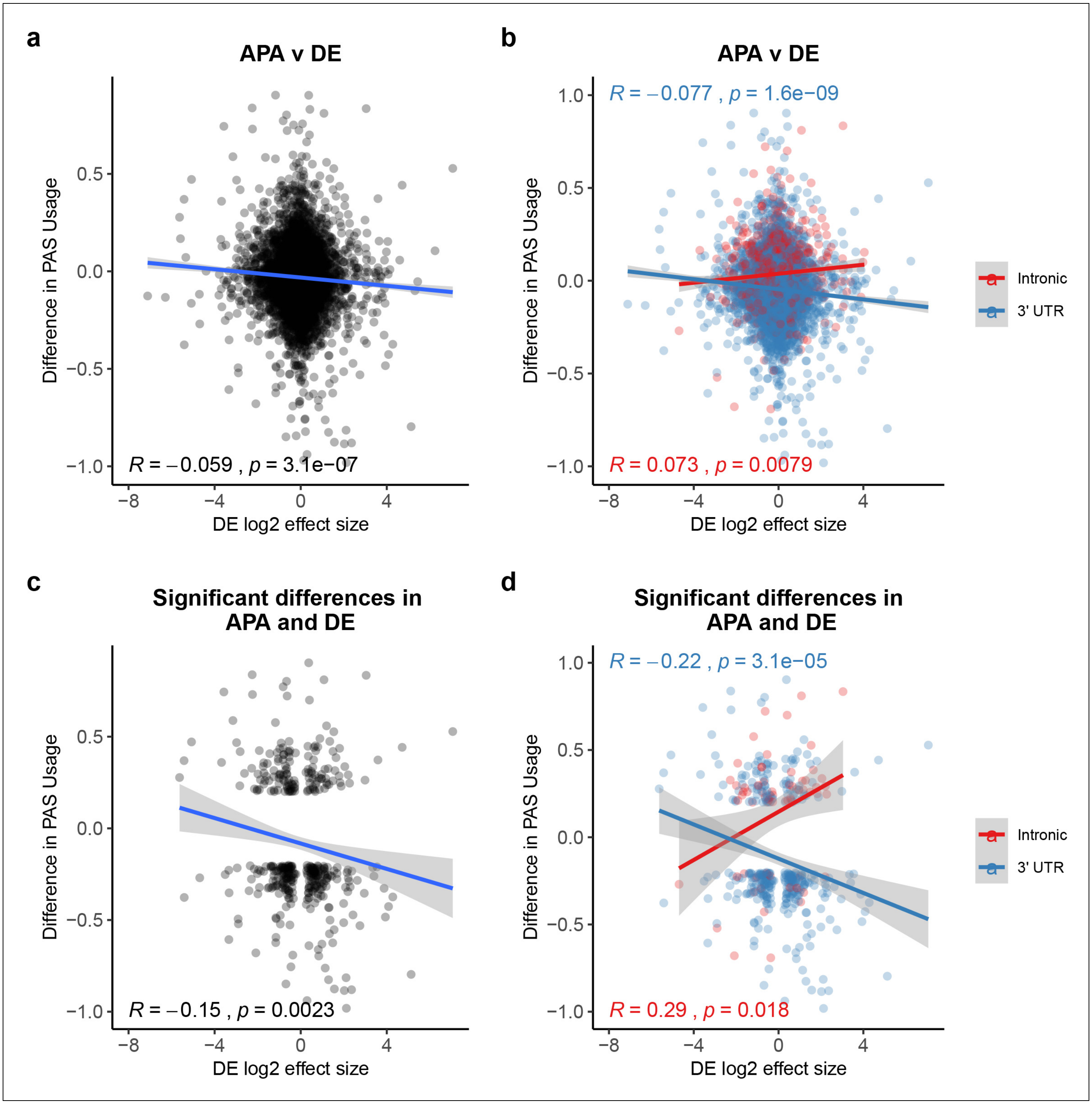

Figure 3. Polyadenylation sites (PAS) usage differences for intronic and 3′ UTR PAS correlate with differential (DE) effect sizes at similar magnitudes but in opposite directions. (**a**) Changes in polyadenylation site usage (ΔPAU) for top intronic or 3′ UTR PAS per gene (see 'Materials and methods') plotted against DE effect size from differential expression analysis. Pearson's correlation plotted, Spearman's correlation R = −0.053. (**b**) ΔPAU for top intronic or 3′ UTR PAS per gene (see 'Materials and methods') plotted against DE effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. Pearson's correlation plotted, Spearman's correlation intronic R = 0.046, Spearman's correlation 3′ UTR R = −0.054. (**c**) ΔPAU for top intronic or 3′ UTR PAS per gene (see 'Materials and methods') plotted against DE effect size from differential expression analysis. Pearson's correlation plotted, Spearman's correlation R = −0.16. (**d**) ΔPAU for top intronic or 3′ UTR PAS per gene (see 'Materials and methods') plotted against DE effect size from differential expression analysis for genes with significant differences in each phenotype at 5% FDR. Pearson's correlation plotted, Spearman's correlation intronic R = 0.22, Spearman's correlation 3′ UTR R = −0.22. In all panels, we calculated the linear

Figure 3 continued on next page

Figure 3 continued

regression. In all panels, negative ΔPAU and DE effect sizes represent upregulation in chimpanzees. In panels **b** and **d**, we colored the points and regressions by genic location.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. ***Figure 3*** relationships expanded to total usage.
Figure supplement 2. Differentially used 3′ UTR polyadenylation sites (PAS) have higher AU element content.
Figure supplement 3. ***Figure 3*** without genes affected by liftover.
Figure supplement 4. Differential expression quality control plots.

variation by summarizing the PAS diversity for a given gene using a single statistic, rather than by analyzing the usage of each site separately. To do so, we measured isoform diversity using Simpson's D (D), a metric traditionally employed by ecologists to measure taxon diversity between environments (***Morris et al., 2014***). In our system, higher D values indicate that the usage is spread more evenly across all PAS for a gene, while low D values suggest the one PAS is more dominant than others (see 'Materials and methods'). As expected, in both humans and chimpanzees, D values are correlated with the number of PAS per gene (***Figure 2—figure supplements 4***, ***5*** and ***6***; human Pearson's correlation 0.62, $p<2.2\times10^{-16}$; chimpanzee Pearson's correlation 0.63, $p<2.2\times10^{-16}$).

Using Simpson's D values calculated for each gene in each individual, we identified (at 5% FDR) 881 genes with significant differences in isoform diversity between species (***Figure 2A***, see 'Materials and methods'). Of these, 426 are genes for which we did not previously detect an inter-species difference in PAS usage, indicating that Simpson's D is capturing an additional dimension of, or is more sensitive to, APA variation between species (***Figure 2—figure supplement 7***; for example, see ***Figure 5—figure supplement 1***).

We proceeded by focusing on genes with low isoform diversity, suggesting a single dominant PAS. We calculated a dominance metric for each gene as the difference in mean usage between the first and second most used PAS (we used different cutoffs to classify dominance; see 'Materials and methods'). We found that the classification dominant PAS is highly consistent across species, a result that is quite robust with respect to the approach used to classify PAS as dominant (***Figure 2B and C***). While the dominant PAS is the same for most genes in humans and chimpanzees, differences in the usage of a dominant PAS are likely to contribute more to differential APA that have functional consequences between species than differences in other PAS. Indeed, regardless of the specific cutoff we used to define dominant PAS, when the dominant PAS is not the same in humans and chimpanzees, the corresponding genes are more likely to be differentially expressed between the species compared with genes where the dominant PAS is the same in both species, (for cutoffs between 0.2 and 0.7, all [$p<0.005$]), and even compared with genes in which only a non-dominant PAS is differentially used ($p>0.8$ for all cutoffs; ***Figure 4A,B***).

In a previous study that collected mRNA from a larger panel of human, chimpanzee, and rhesus macaque LCLs, Khan et al. identified genes whose regulation likely evolves under directional selection in humans and chimpanzees (***Khan et al., 2013***). We were able to consider RNA and protein expression data as well as APA data from 2532 genes. We found that 22 of the genes with significant inter-species differences in APA at both the site level and in isoform diversity are among those whose regulation likely evolves under directional selection in the chimpanzee lineage, a 1.6× enrichment over what is expected by chance (hypergeometric test, $p=0.015$). While we did not identify any significant gene ontology (GO) categories for these genes, 5 of the 22 genes are associated with protein transport (CAPZA1, TPM3, TMED2, GOSR2, AP3S1) and 2 are ribosomal subunits (RPL13 and RPL7L1). We did not find a similar enrichment when we considered genes whose regulation evolved under selection in humans, but the sample size is rather small.

Variation in APA and differences in protein expression

Given the well-characterized molecular connection between APA and the regulation of protein translation, we hypothesized that genes with inter-species differences in APA are also more likely to be differentially translated between the species (***Di Giammartino et al., 2011***; ***Floor and Doudna, 2016***; ***Tian and Manley, 2017***). To examine this, we obtained estimates of protein translation based on ribosome profiling data that were collected from human and chimpanzee LCLs by ***Wang et al.,***

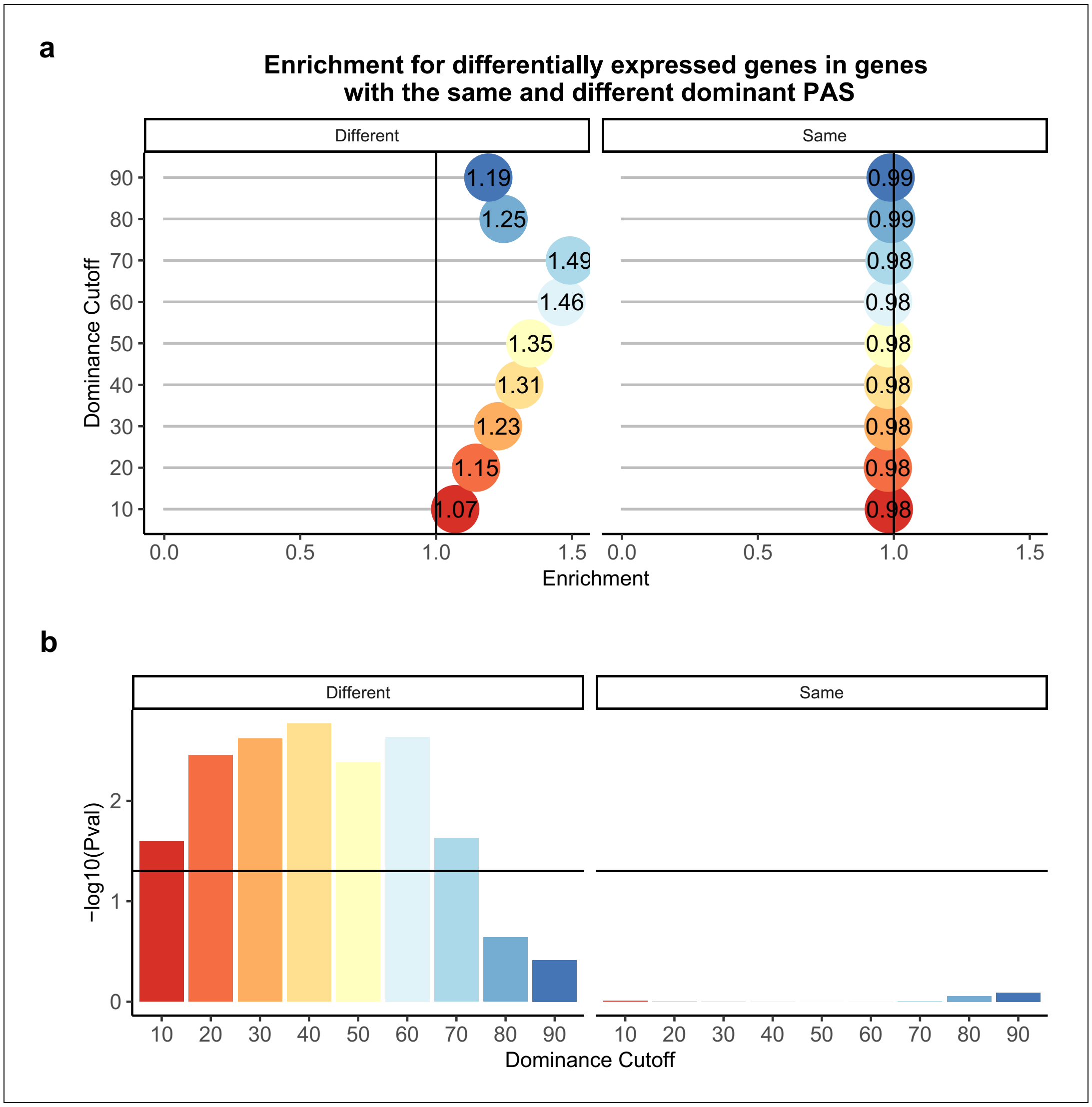

Figure 4. Difference in dominant polyadenylation sites (PAS) between species likely drives differences in expression. (**a**) Enrichment of genes with the different (left) of same (right) dominant PAS by dominant cutoff in differentially expressed genes. (**b**) $-\log_{10}$ (p-values) for enrichments in (a) calculated with hypergeometric tests. Horizontal line represents a p-value of 0.05. The bars are colored by the dominance cutoff on the X axis.
The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. ***Figure 4*** without genes affected by liftover.

2018b. At a 5% FDR, Wang et al. identified 1993 differentially translated genes between humans and chimpanzees. Genes with significant inter-species differences in isoform diversity, but without significant differences in the usage at individual PAS, are enriched among the differentially translated gene set (1.21×, hypergeometric test, p=0.011; ***Figure 5A,B***). The genes with differentially used PAS are 32× enriched within the differentially translated genes for genes involved in translation initiation (hypergeometric test, FDR 0.0091, ***Supplementary file 1***, see 'Materials and methods').

We next investigated the relationship between ΔPAU in humans and chimpanzees and the effect sizes for differences in protein translation between the species (***Wang et al., 2018b***). Considering the most differentially used 3' UTR or intronic PAS per gene (see 'Materials and methods'), we identified a significant correlation between inter-species differences in translation and ΔPAU for 3' UTR PAS, with a stronger correlation among genes with significant differences in both APA and translation (***Figure 5—figure supplement 2***). As expected, and to some extent we view this as a control

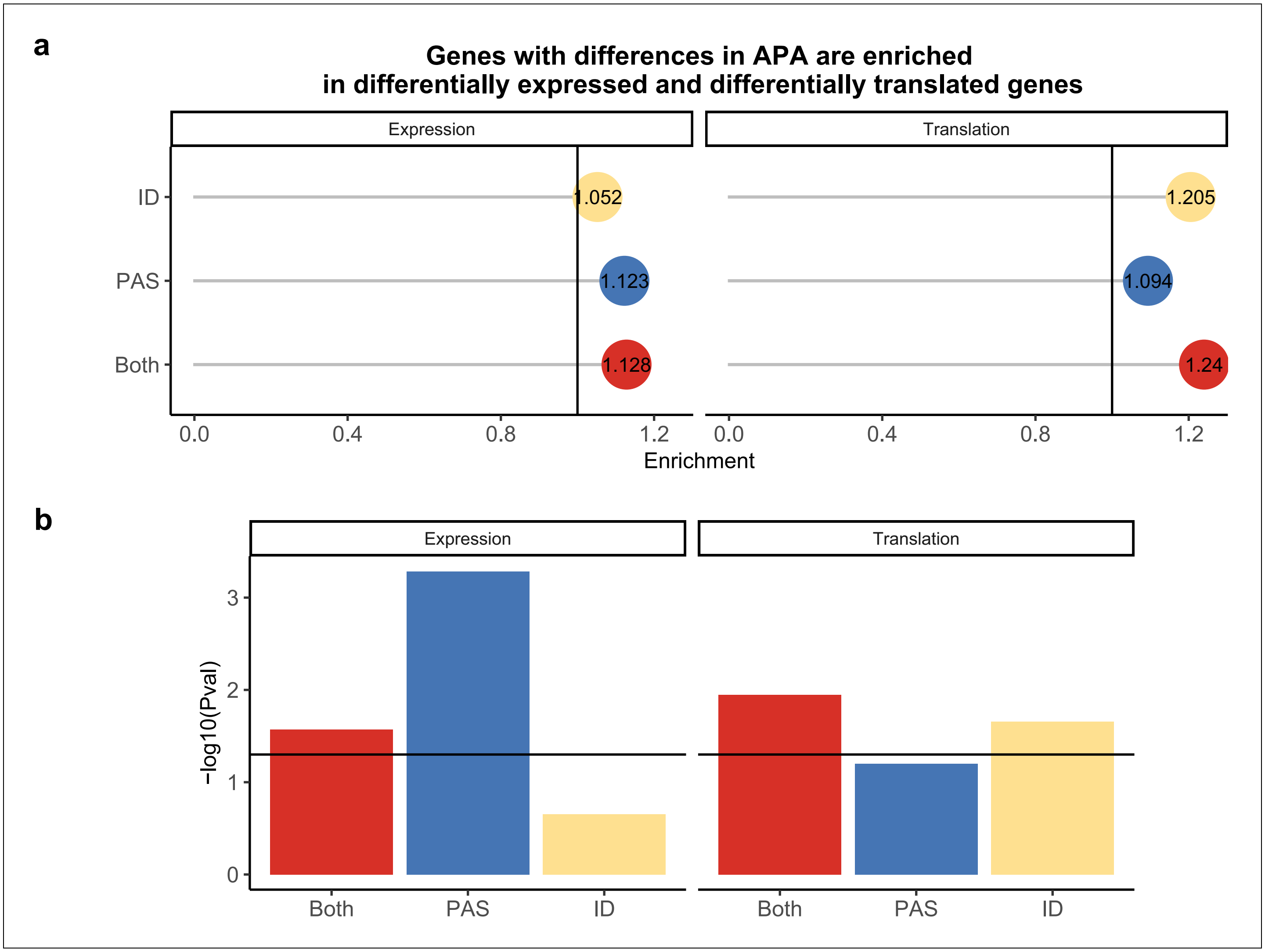

Figure 5. Polyadenylation sites (PAS)-level differences in alternative polyadenylation (APA) may drive differences in expression while isoform diversity differences likely drive translation differences. (a) Enrichment of genes with isoform-level differences (ID), differences in APA at PAS level (PAS), or at both levels (Both) within differential expressed genes and differentially translated genes. Differentially translated genes reported by ***Wang et al., 2018b***. (b) $-\log_{10}$ (p-values) for enrichments in A calculated with hypergeometric tests. Horizontal line represents a p-value of 0.05.
The online version of this article includes the following figure supplement(s) for figure 5:

Figure supplement 1. Gene with significant differences in isoform diversity only.
Figure supplement 2. Relationship between changes in polyadenylation site usage (ΔPAU) and differential translation (TE) effect sizes.
Figure supplement 3. Relationship between alternative polyadenylation (APA) differences and protein decay mark.
Figure supplement 4. ***Figure 5*** without genes affected by liftover.

analysis, we did not identify a significant correlation between intronic PAS ΔPAU values and differences in translation (***Figure 5—figure supplement 2***).

Given the apparent impact of PAS usage on protein translation, we next considered direct measurements of protein expression data from 3391 genes in LCLs from humans and chimpanzees (***Khan et al., 2013***). Using summary statistics from this study, we found 1263 genes to be differentially expressed at the protein level between the species (FDR of 5%). As the protein measurements are restricted to these 3391 genes, we do not have enough power to ask if genes with inter-species differences in APA are also more likely to be differentially expressed at the protein level. However, we did find a positive correlation between the absolute value of 3′ UTR ΔPAU and the standardized number of ubiquitination sites for the same gene (Pearson's correlation, R = 0.15, $p=5.0\times10^{-7}$,

Figure 5—figure supplement 3, see 'Materials and methods'), consistent with the observation that 3' UTR PAS are targets for the regulation of protein decay (***Dubnikov et al., 2017***; ***Ravid and Hochstrasser, 2008***). Thus, we next focused on the 506 genes with significant inter-species differences in protein expression and an absence of corresponding differences in transcript expression levels that we also tested for differences in APA. Khan et al. reasonably hypothesized that inter-species differences in translation could account for the emergence of differences in protein expression levels when there are no regulatory differences at the RNA level, but they were unable to point to specific mechanisms. These genes are particularly interesting in the context of our current study, because APA which results in changes to 3' UTR length may be more likely to result in differences in protein expression without affecting the expression level of the mRNA.

Indeed, we found 76 genes with inter-species differences in APA that are also differentially expressed at the protein but not at the RNA level between humans and chimpanzees (***Figure 6A,B***). In these 76 genes, inter-species differences in PAS usage are enriched at the 3' UTR (***Figure 6—figure supplement 1***). The 76 genes are likely of functional relevance, compared to all of the genes, with at least one differentially used PAS being enriched for cellular components such as the protease complex and endopeptidase complex. The set is also enriched for processes such as the regulation of DNA-templated transcription and amino acid activation. (For a full list, see ***Supplementary file 1***, see 'Materials and methods'.) Finally, to assess whether APA contributes to differences in gene regulation by affecting translation efficiency or protein degradation, we asked whether genes with differential protein expression were also differentially translated. Of the 149 genes with significant differences in APA and protein expression, Wang et al. reported translation measurements for 142 (***Wang et al., 2018b***). Only 34 genes displayed significant differences in translation efficiency, suggesting that isoform-specific post-translational modification of protein levels is largely responsible for protein-level differences (***Figure 6C,D***).

Discussion

Comparative primate functional genomic studies have contributed to our understanding of the gene regulatory processes that underlie genotype-phenotype relationships. A common goal of these studies is to understand the general properties and level of conservation of specific regulatory phenotypes, such as gene expression or DNA methylation levels. Multiple data types collected from the same cell lines or tissues can then be analyzed together to generate hypotheses about how gene regulatory processes contribute to inter-species differences in morphology, physiology, cognitive phenotypes, disease susceptibility, and other traits (***Blake et al., 2020***; ***Khan et al., 2013***; ***Romero et al., 2018***; ***Wang et al., 2018b***; ***Zhou et al., 2014***). Moreover, unlike functional genomic studies within humans, comparative studies require only modest sample sizes to identify regulatory effects. This is because genetic variation between species is greater than genetic variation within species. Thus, regulatory differences that distinguish humans from other primates tend to have larger effect sizes than regulatory differences that distinguish between individual humans (***Housman and Gilad, 2020***).

Importantly, inter-species differences identified using a comparative approach may be important not only for understanding primate evolution but may implicate candidate loci for further investigation in humans. For example, genomic regions that are conserved in primates may point to loci that are likely to have negative functional consequences in humans, potentially with effects on disease risk (***Housman et al., 2019***). Identification of genomic regions under adaptation in humans is also critical, as they may point to causal mechanisms for human-specific traits, including diseases that are specific to humans (***Gokhman et al., 2020***; ***Ward and Gilad, 2019***).

We characterized APA in human and chimpanzee LCLs to begin to understand the role that co-transcriptional mechanisms play in the evolution of gene regulation in primates. Our group has previously studied a variety of other gene regulatory phenotypes in primate LCLs (***Cain et al., 2011***; ***Khan et al., 2013***; ***Wang et al., 2018b***; ***Zhou et al., 2014***). We and others have demonstrated that gene regulatory phenotypes in these cell lines recapitulate many regulatory patterns seen in primary tissues (***Caliskan et al., 2011***; ***Çalışkan et al., 2014***; ***Khaitovich et al., 2006***). Not only did our use of primate LCLs allow us to circumvent many of the practical and ethical issues associated with primate research, but it also allowed us to integrate 3' Seq and gene expression data from this study in the context of previously collected ribosome occupancy and protein expression data from primate

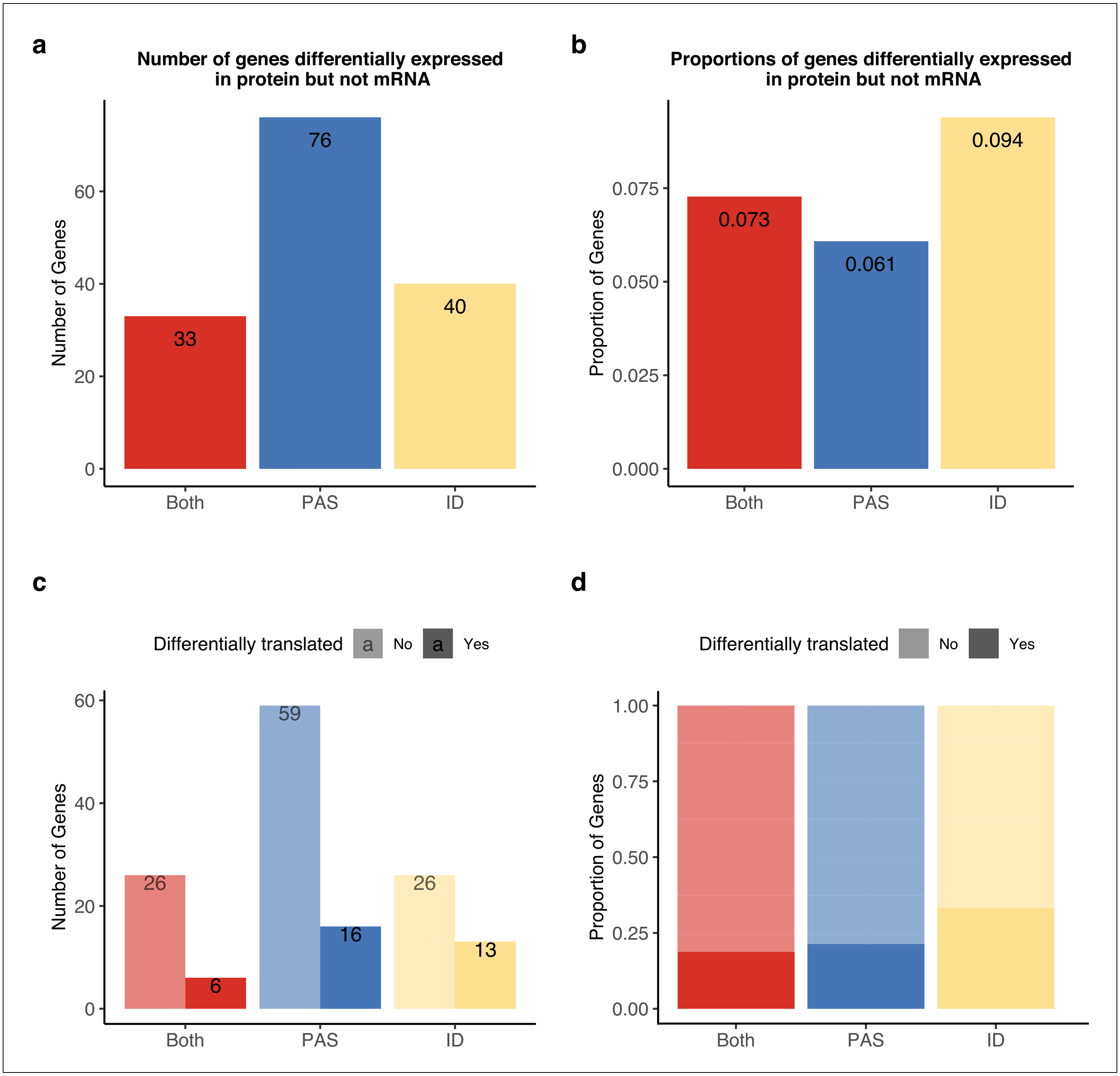

Figure 6. Alternative polyadenylation (APA) differences explain genes differentially expressed at protein level but not in mRNA. APA likely mediates functional differences post-translationally. (**a**) Number of genes with isoform-level differences (ID), differences in APA at polyadenylation sites (PAS) level (PAS), or at both levels (Both) differentially expressed in protein (5% FDR) but not mRNA (5% FDR). Genes with differentially expressed protein reported in ***Khan et al., 2013***. (**b**) Proportion of genes with differential usage at PAS level (1251 genes), isoform diversity level (426 genes), or both (454 genes) differentially expressed in protein (5% FDR) but not mRNA (5% FDR). (**c**) Genes reported in (a) separated by genes differentially translated at 5% FDR. Differentially translated genes reported in ***Wang et al., 2018b***. (**d**) Genes differentially expressed in protein but not in mRNA, colored by differences in APA. Proportion of genes in the set differentially translated at 5% FDR.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Enrichment of 3′ UTR polyadenylation sites (PAS) in genes differentially expressed in protein but not in mRNA.

Figure supplement 2. ***Figure 6*** without genes affected by liftover.

LCLs (**Khan et al., 2013**; **Wang et al., 2018b**). Together, these data allowed us to study the contribution of APA to inter-species differences in transcript and protein expression levels. We recognize the limitations of this study with respect to physiologically interesting phenotypes. The genome-wide map of APA events in human and chimpanzee has allowed us to infer global mechanisms connecting APA to other gene regulatory processes. By expanding the study of primate APA into other cell types and dynamic processes, future studies will be able to connect the mechanisms described here to primate phenotypes of interest.

APA is an important molecular mechanism with regard to both the evolution of gene regulation and physiological traits. On a long-term evolutionary scale, both 3' UTR length and the proportion of genes exhibiting APA have scaled with genome size and complexity. The expansion of APA is believed to have introduced biological complexity independent of an increase in the number of distinct genes (**Mayr, 2016**; **Mayr, 2017**). As usage of multiple isoforms has been maintained, it is likely that distinct isoforms have divergent functions that are maintained by balancing selection. For example, APA facilitates post-transcriptional regulation of a *Drosophila Hox* gene through maintenance of two isoforms differentially targeted by multiple miRNAs (**Patraquim et al., 2011**). In turn, genome-wide changes in APA during differentiation of stem cells to terminal cell types direct isoform-specific gene regulation that is important for development in a range of species, including humans (**Hilgers et al., 2011**; **Ji et al., 2009**; **Li et al., 2012**). Further, dysregulation of tumor suppressor genes through intronic polyadenylation is known to contribute to cancer pathogenesis (**Dubbury et al., 2018**; **Lee et al., 2018**). We hypothesize that a better understanding of APA in primates will aid in the understanding of APA evolution and its contribution to human-specific phenotypes.

APA is mostly conserved, especially dominant sites

We measured APA from 3' Seq data by calculating a ratio of isoforms terminating at one PAS compared to isoforms terminating at other PAS for the same gene. We then compared PAS usage ratios between species. To expand our understanding of APA conservation, we also calculated an isoform diversity statistic (Simpson's D) for each gene in each species. Because Simpson's D captures both the number of PAS isoforms and their usage, we were able to evaluate small regulatory changes spread across many PAS, rather than only focus on large changes at individual PAS. While previous studies have used Shannon index to quantify isoform diversity (**Pai et al., 2016**; **Wang et al., 2018b**), we found Simpson's D to be less correlated with PAS number, making it less sensitive to the number of PAS per gene (**Figure 2—figure supplements 5** and **6**). In addition, by placing more weight on dominant PAS, Simpson's D more closely mirrors our current biological understanding of APA, wherein dominant PAS play a larger role in downstream gene regulation (**Morris et al., 2014**).

In general, we found that both individual PAS usage and isoform diversity are highly conserved between human and chimpanzee. Consistent with comparative studies of APA in humans and rodents, which used genomic synteny to identify conserved PAS, we found higher conservation among genes with a single PAS (**Ara et al., 2006**; **Wang et al., 2018a**) and showed that sequence variation in PAS signal sites and the surrounding U-rich regions contributes to inter-species differences in APA (**Wang et al., 2018a**). Because we characterized APA in closely related primates, our study provides additional insight into APA divergence at both the gene level and the species level, revealing functional changes that contribute to differences in downstream gene regulation. For example, we observed that when genes use one PAS markedly more often than others, said dominant PAS tends to be the same in both human and chimpanzee. It is likely that strong selection pressures have acted on these genes, resulting in continual usage of the same dominant isoform. This could imply that the dominant isoform is functionally important and alternative isoforms are potentially associated with reduced fitness. However, non-dominant isoforms also show evidence of conservation. Thus, it remains possible that there is a threshold at which the level of expression of the alternative isoforms begins to impede gene function.

APA is associated with gene expression divergence

Our study also revealed that the majority of differentially used PAS between species are located in 3' UTRs. We showed that, across species, increased intronic PAS usage is associated with a modest increased mRNA expression levels, while increased 3' UTR PAS usage is correlated with a modest decrease in mRNA expression. In a previous study, we found that human alternative polyadenylation quantitative trait loci (apaQTL) alleles associated with increased intronic PAS usage were correlated with *decreased* mRNA expression levels (**Mittleman et al., 2020**). This is not the first molecular phenotype wherein a within-species study revealed alternative regulatory models compared to an inter-species analysis. For example, Pai et al. reported tissue-specific differential methylation to be almost exclusively inversely correlated with gene expression patterns between human and chimpanzee (**Enard et al., 2004**; **Pai et al., 2011**; **Weber et al., 2007**), whereas Banovich et al. discovered genetic variation associated with DNA methylation variation that was both directly and inversely correlated with expression quantitative trait loci (eQTLs). At this time, we cannot provide evidence for a mechanistic explanation for these contradictory observations. We hypothesize the following: Transcripts terminating in introns are likely subject to nonstop decay (NSD). By studying APA variation within humans, we probably captured the effects of intronic termination. Across species, however, increased intronic PAS may simply track the overall expression level of the gene. Specifically, increased usage of intronic PAS may result in truncated isoforms that do not contain 3′ UTR *cis* regulatory elements that would normally signal mRNA decay. If these truncated isoforms are no longer targets of mRNA decay, this could cause some of these genes to appear upregulated. We hypothesize that the within-species effects related to NSD were likely overshadowed by inter-species differences, which typically have much larger effect sizes than differences observed within a population (**Housman and Gilad, 2020**).

Alternatively, the large effect sizes for differential usage of 3′ UTR PAS could also be driving the relationship between differential APA and differential expression. In line with this hypothesis, we found increases in AU destabilizing elements and ubiquitination marks for genes with divergent 3′ UTR PAS. However, since PAS usage is calculated as a ratio, we may have detected changes in intronic PAS usage solely as a mathematical consequence of changes in 3′ UTR PAS. Functional follow-up on the genes with PAS detected as differentially used between and within species would be necessary to explore the relative importance of each of these regulatory pathways and to disentangle the results from both studies.

In past studies, we and others have estimated the proportion of variation in gene expression explained by different regulatory mechanism. For example, we have previously tested for differences in expression before and after accounting for another regulatory mechanism or used formal mediation analyses (**Blake et al., 2020**; **Blekhman et al., 2009**; **Cain et al., 2011**; **Eres et al., 2019**). We would have liked to perform similar analysis in the current study, regarding the role that APA plays in the overall regulatory divergence between the species. However, APA was measured using ratios of alternative mRNA isoforms and thus, effect sizes for APA and differential expression are on different scales and we cannot use a standard mediation approach to formally calculate the proportion of expression variation explained by APA. That said, we are generally convinced that APA contributes to differences in mRNA expression overall because 764 of 3796 (20.1%) differentially expressed genes also have significant differences in APA.

Inter-species differences in APA explain protein-specific regulatory divergence

Though genes are ultimately expressed as proteins, many studies (including current studies by our group) still measure mRNA expression as an implicit proxy for protein expression levels. As a justification for this approach, we typically point to the fact that after accounting for technical considerations, the correlation between mRNA levels and protein abundance is quite high genome-wide, specifically, across genes (**Buccitelli and Selbach, 2020**; **Csárdi et al., 2015**). However, we also know that at the level of a single gene, across individuals or tissues, the correlation between mRNA and protein measurements tends to be much lower (**Battle et al., 2015**; **Buccitelli and Selbach, 2020**). This suggests that a number of molecular mechanisms decouple mRNA and protein expression levels post-transcriptionally. Clearly, we do not yet fully understand the post-transcriptional and translational mechanisms that shape the proteome (**Buccitelli and Selbach, 2020**).

Within human populations and between primates, there is a large number of genes that are differentially expressed at the mRNA level but not as proteins. By directly measuring translation levels for these genes, previous studies have proposed that post-translational protein buffering can explain the decreased variation at the protein level (**Battle et al., 2015**; **Wang et al., 2018b**). Conversely, there are also genes that are more variable at the protein levels than at the mRNA levels

(**Battle et al., 2015**; **Chick et al., 2016**; **Khan et al., 2013**). Our previous work demonstrated that some protein-specific QTLs are also highly correlated with differences in APA (**Mittleman et al., 2020**). Considered all of these observations, we expanded this analysis and demonstrated that genes that are differentially expressed as proteins, but not at the mRNA level, between human and chimpanzee, tend to have divergent APA patterns. We also found that the divergent protein levels are likely due to post-translational molecular mechanisms. While we cannot directly test the mechanism here, we hypothesize that APA could lead to variation in protein levels as a consequence of protein autoregulation, by differentially including RNA and protein-binding motifs (**Buccitelli and Selbach, 2020**; **de Bie and Ciechanover, 2011**; **Müller-McNicoll et al., 2019**). Alternatively, APA could contribute to temporal and spatial differences in protein expression, which would affect our ability to quantify protein with traditional techniques (**Buccitelli and Selbach, 2020**; **Tian and Manley, 2017**).

In conclusion, a better understanding of co-transcriptional gene regulatory mechanisms, such as APA, may point to additional mechanisms contributing to the decoupling of mRNA and protein abundance and more generally, enhance our understanding of how variation percolates through genetic variants, mRNA, and protein to ultimately affect human phenotypic diversity.

Materials and methods

Cell culture and collections

We grew six human and six chimpanzee Epstein-Barr virus-transformed LCLs in glutamine-depleted RPMI (RPMI 1640 1× from Corning [15–040 CM]), completed with 20% FBS, 2 mM GlutaMax (Gibco [35050–061]), 100 IU/mL penicillin, and 100 µg/mL streptomycin. We cultured all cells at 37˚C at 5% CO_2. We passaged each cell line a minimum of three times, then maintained cells at 1×10^6 cells/mL in preparation for collection. Cell line numbers and details can be found in **Supplementary file 4**. The human lines were derived from Yoruba individuals collected as part of the HapMap project and can be ordered through the Coriell Institute (**International HapMap Consortium, 2005**). The sample IDs and Research Resource Identifiers for the human cell lines are the following: NA18498: CVCL_P466, NA18499: CVCL_P457, NA18502: CVCL_P459, NA18504: CVCL_P460, NA18510: CVCL_P461, NA18523: CVCL_P468. Chimpanzee LCLs were originally transformed from individuals from the New Iberia Research Center (University of Louisiana at Lafayette), Coriell IPBR repository and Arizona State University (**Khan et al., 2013**). The cell lines have previously been used for similar studies of primate gene regulation (**Cain et al., 2011**; **Khan et al., 2013**; **Wang et al., 2018b**; **Zhou et al., 2014**). Cell lines have been authenticated with RNA-seq and have tested negative for mycoplasm.

Once all cells lines reached 1×10^6 cells/mL, we used the collection and RNA extraction method detailed in **Mittleman et al., 2020** . to extract whole-cell and nuclear mRNA. Briefly, we collected 30 million cells in two 15 million cell aliquots. We extracted nuclei from one aliquot per line using the nuclear isolation protocol outlined by **Mayer and Churchman, 2016**. We extracted mRNA in two fraction- and species-matched batches, using the miRNeasy kit (Qiagen) according to manufacturer's instructions, including the DNase step to remove genomic DNA. We quantified mRNA and tested quality using a nanodrop. Details of mRNA processing for each line, including concentrations and quality, can be found in **Supplementary file 4**.

3' Seq to identify PAS and quantify site usage

We generated 3' Seq libraries from whole-cell and nuclear-isolated mRNA from six chimpanzee and six human individuals using the QuantSeq Rev 3' mRNA-Seq Library Prep Kit (**Moll et al., 2014**) according to the manufacturer's instructions. We sequenced all libraries on the Illumina NextSeq500 at the University of Chicago Genomics Core facility using single-end 50 bp sequencing.

We mapped human 3' Seq libraries to GRCh38 (**Schneider et al., 2017**) and chimpanzee libraries to panTro6 (**Chimpanzee Sequencing and Analysis Consortium, 2005**) using the STAR RNA-seq aligner with default settings (**Dobin et al., 2013**). Similar to our previous work, we removed reads with evidence of internal priming resulting from the poly(dT) primer. We filtered reads proceeded by 6 As or 7 of 10 As in the base pairs directly upstream of the mapped location (**Mittleman et al., 2020**; **Sheppard et al., 2013**; **Tian et al., 2005**). To ensure that differences in low quality bases would not bias our results, we treated any N in the genome annotation as an A. All raw read counts, mapped read counts, and filtered read counts can be found in ***Supplementary file 4***.

We first identified an inclusive set of PAS in each species separately. We used the same in-house peak caller described in ***Mittleman et al., 2020***, annotating each PAS as the most 3′ base in each peak. The initial PAS set included 340,023 in human and 303,249 in chimps. We extended PAS 100 bp upstream and 100 bp downstream and used a reciprocal liftover pipeline to identify an inclusive set of orthologous PAS. We downloaded chain files from UCSC Genome Browser (***Kent et al., 2002***). Details of the pipeline and number of PAS passing each step can be found in ***Figure 1—figure supplement 12***.

Due to gene annotation differences between species, we annotated all orthologous PAS to the human NCBI RefSeq annotation downloaded from UCSC Genome Browser (***Pruitt et al., 2004***). We used a hierarchical model to assign PAS to genic locations (***Lin et al., 2012***; ***Mittleman et al., 2020***). We prioritized annotations in the following order: 3′ UTRs (UTR5), 5 kb downstream of genes (end), exons (cds), 5′ UTRs (UTR5), and introns (intron). We quantified reads mapping to each annotated PAS for each individual in both the total RNA libraries and nuclear RNA libraries using featureCounts with the -s strand specificity flag (***Liao et al., 2014***). We calculated usage for each PAS in each library as a ratio of reads mapping to the PAS divided by the number of reads mapping to any PAS in the same gene (***Figure 1—figure supplement 2***). We implemented two filtering steps to remove PAS with ratios likely biased by low site count or low gene count separately in each fraction.

Next, we filtered out sites with less than 5% usage in both species in the nuclear fraction. We then merged nuclear counts across all PAS in each gene. We removed PAS in genes not passing a cutoff of $\log_2$(CPM) >2 in at least 8 of the 12 individuals. After applying these filters, we were left with 44,432 PAS. As a quality control metric, we compared PAS usage calculated from the nuclear fraction to PAS usage calculated from whole-cell fraction for each individual (we used the same methods to identify and quantify PAS usage in the whole-cell 3′ Seq data). We expected a high correlation between PAS usage in each fraction. Further, we expected a similar correlation in human and chimpanzee individuals (***Mittleman et al., 2020***). Human individual NA18499 had significantly lower across-species correlation than the other individuals and was therefore removed from the analysis (***Figure 1—figure supplement 3***).

To ensure gene expression level did not introduce ascertainment bias, we tested the relationship between PAS number and normalized gene expression. In both species, the number of PAS is negatively correlated with normalized gene expression (human: Pearson's correlation = -0.19, $p<2.2\times10^{-16}$; chimpanzee: Pearson's correlation = -0.17, $p<2.2\times10^{-16}$; ***Figure 1—figure supplement 6***). We expected species to contribute the most amount of variation to PAS usage. We ran principal component analysis (PCA) on the filtered nuclear PAS usage. PC1 accounts for 41.8% of the variation and is highly correlated with species (R^2 = 0.68). PC2 accounts for 13.1% of the variation and is moderately correlated with RNA extraction technician (R^2 = 0.38) and extraction day (R^2 = 0.28). As both of these variables are balanced with respect to species, we do not believe they bias the results (***Figure 1—figure supplement 5***). We identified 302 sites used at a rate of 5% or greater in humans and 0% in chimpanzees, which we designated as human-specific. We identified 357 sites used at a rate of 5% or greater in chimpanzee and 0% in humans, which we designated as chimp-specific.

We acknowledge the possibility that unlifted PAS may affect the downstream analyses; therefore, we removed genes for which PAS ratios may be affected. Specifically, we annotated and calculated usage for the human PAS, including the 10,077 PAS that do not reciprocally lift to the chimpanzee genome. After removing PAS in genes previously identified as lowly expressed and PAS with usage below 5%, 386 PAS in 353 genes remain (***Figure 2—figure supplement 8***). We removed these 353 genes and recreated main figures 3-6. (***Figures 3–6***, ***Figure 3—figure supplement 3***, ***Figure 4—figure supplement 1***, ***Figure 5—figure supplement 4***, ***Figure 6—figure supplement 2***).

Orthologous 3′ UTRs

We identified a set of orthologous UTRs using the orthologous exon file described in the differential expression analysis section of the 'Materials and methods'. We merged all regions annotated as 3′ UTR by gene. If a gene had multiple non-continuous annotations, we selected the most 3′ region as the orthologous UTR. We used deepTools compute matrix and plotHeatmap functions to plot merged human and chimpanzee reads along the orthologous 3′ UTR set (***Figure 1—figure***

supplement 1; ***Ramírez et al., 2016***). For all genes with PAS only in 3′ UTRs, we assigned PAS to single, first, middle, and last, as previously described (***Wang et al., 2018a***).

Analysis of sequence conservation around PAS

We used PhyloP scores to measure sequence-level conservation. We downloaded the hg38 100-way vertebrate PhyloP bigwig file from the UCSC table browser (***Pollard et al., 2010***). We computed scores for PAS regions as well as 200 bp intervals by taking the mean of the base pair scores. We removed any region with missing data from the analysis. We tested for differences in mean phloP scores using Wilcoxon rank sum tests.

We tested for the presence of the polyadenylation signal site motif in the 200 bp PAS regions. We used the bedtools nuc tool with the strand-specific flag to test for the presence of each of the 12 previously annotated motifs for each PAS in both species (***Beaudoing et al., 2000***; ***Quinlan and Hall, 2010***). If a PAS had multiple motifs, we used a hierarchical model to choose the site based on the number of PAS with each identified motif (order: AATAAA, ATTAAA, AAAAAG, AAAAAA, TATAAA, AATATA, AGTAAA, AATACA, GATAAA, AATAGA, CATAAA, ACTAAA). The proportion of PAS with each signal site motif matched across species (***Figure 2***). To ask if the presence or absence of a signal site explained species specificity or site-level differences, we restricted our analysis to the top two signal sites. These two motifs are the only sites where the presence of a signal is associated with increased usage of the site in both species (***Figure 1—figure supplement 10***). For the 359 PAS with one of these two signal sites present only in chimpanzees, average usage was higher in chimpanzees than in humans (p=0.025). For the 361 PAS with one of these two signal sites present only in humans, average usage was higher in humans ($p=2.0\times10^{-4}$). We used hypergeometric tests to evaluate enrichment of differentially used PAS and species-specific PAS in the set of PAS with signal sites in only one species.

We also examined the proportion of U nucleotides in each PAS region. We used the bedtools nuc with the -s flag for strand specificity (***Quinlan and Hall, 2010***). We tested if PAS with differences in U content are enriched for differentially used PAS using a hypergeometric test.

Differential APA

PAS-level differences

We quantified reads mapping to each PAS using the featureCounts tool with the -s strand specificity tool (***Liao et al., 2014***). We tested for site-level differences between human and chimpanzee using the leafcutter leafcutter_ds.R tool with standard settings (***Li et al., 2018***). We tested for differences in both the total and nuclear fractions. We tested 43,038 PAS in 8422 genes in the nuclear fraction and 41,914 PAS in 8333 genes in the total fraction. We classified PAS as differentially used if the gene reached significance at 5% FDR and the PAS had a ΔPAU greater than 20% (absolute value [ΔPAU]>0.2). A negative ΔPAU indicates increased usage in chimpanzees and ΔPAU indicates increased usage in humans. The top PAS per gene is the PAS with the most significant difference between species; ties were broken using mean usage for all individuals in both species.

Isoform diversity differences

We calculated Shannon information content $(-\sum_{i=1}^{s} p_i \log_2 P_i)$ and Simpson index $(1-\sum_{i=1}^{s} p_i^2)$ using mean usage of each PAS in humans and chimpanzees, where p_i^2 is the usage of the ith of *s* sites in the gene. We used Simpson index to assess isoform diversity because the correlation between Simpson index and number of PAS is lower than the correlation between Shannon information content and the number of PAS per gene (***Figure 2—figure supplements 5***, ***6***). To identify genes with differences in isoform diversity, we recalculated Simpson index per gene per individual and tested for differences between species with Wilcoxon tests. We reported genes with differences at 5% FDR.

Conservation of dominant PAS

We consider a gene to have a dominant PAS if the within species average usage of the top used PAS is greater than the second most used site by 0.4. We reported results for cutoffs between 0.1 and 0.9. If a gene had a dominant PAS in either species, we included the top used site for both species when testing if genes use the same or different dominant PAS between species. We tested for enrichment of genes using the same or different dominant PAS with differentially expressed genes using hypergeometric tests.

Differential expression analysis

We generated unstranded RNA-seq libraries using the Illumina TruSeq Total RNA kit according to the manufacturer's instructions using the total mRNA collected from all 12 individuals (Illumina, San Diego, CA, USA). We sequenced RNA-seq libraries at the University of Chicago Genomics Core facility using the single-end 50 bp protocol on one lane of the Illumina HiSeq 4000 machine. RNA quality and concentration at the time of library prep and number of sequenced reads per library are available in ***Supplementary file 5***. We mapped the human libraries to GRhg38 (***Schneider et al., 2017***) and chimpanzee libraries to panTro6 (***Chimpanzee Sequencing and Analysis Consortium, 2005***) and quantified reads mapping to orthologous exons.

To generate an updated orthologous exon file for the most recent chimpanzee genome assembly (panTro6), we followed the procedure reported in Pavlovic et al. with slight modifications (***Pavlovic et al., 2018***). We started with human (GRCh38) exon definitions from Ensembl version 98. We filtered this set of definitions for biotypes 'protein_coding' using the command mkgtf from cellranger (10× genomics). We then removed exon segments that were in exon definitions for multiple genes. This broke some exons into smaller unique exons. We then removed exons smaller than 10 bp. We took the final set of exons (1,371,917 exons from 20,338 genes) and extracted their sequences from the genome Ensembl GRCh38.p12. We used BLAT version 35 to identify orthologous sequences within the chimpanzee genome (panTro6) (***Kent, 2002***). We removed hits with indels larger than 25 bp (using a function blatOutIndelIdent from https://bitbucket.org/ee_reh_neh/orthoexon). We then extracted the panTro6 sequences that had the highest sequence identity. We ran BLAT on this orthologous exon set to find matches in both the human and chimpanzee genomes. We removed exons that did not return the original location in humans or chimpanzees, as well as exons that mapped to multiple places with higher than 90% sequence identity. We removed exons from different human genes that mapped to overlapping regions in the chimpanzee genome. Finally, we removed exons that mapped to a different contig than the majority of exons from each gene. This resulted in a set of 1,250,820 orthologous exons from 19,515 genes.

We mapped on average 18.6 million reads to orthologous exons. We collapsed orthologous exons to quantify raw gene expression for each gene in each individual. We standardized counts and filtered out genes that did not pass the criteria of $\log_2$(CPM) values greater than 1 in 8 of the 12 individuals. To prepare counts for differential expression modeling, we used the Voom function with the quantile normalization method in the limma R package (***Ritchie et al., 2015***). We used principal component analysis (PCA) to test for batch effects. PC1 explains 35.1% of the variation and is highly correlated with species (R^2 = 0.98) (***Figure 3—figure supplement 4***). Collected metadata such as the percent of live cells at collection, cell concentration at collection, RIN score, and RNA concentration do not segregate by species (***Figure 3—figure supplement 4***). We modeled species as a fixed effect and called genes as differentially expressed at a 5% FDR. The results from our differential expression analysis, including effect sizes and significance values, are available in ***Supplementary file 6***.

Integration of translation and protein data

We downloaded differentially translated genes and their effect sizes from Additional file 5 of ***Wang et al., 2018b***. Wang et al. modeled differential translation using ribosome profiling of four human, four chimpanzee, and four rhesus macaque LCLs. For all integrations, we conditioned on the 6407 genes tested in the Wang et al. study and in our APA analysis. We tested for enrichments using a one-sided hypergeometric test implemented in R. We tested for correlations in effect sizes between site-level ΔPAU and translation HvC effect sizes by first filtering for the top PAS (see top PAS method above, ***Figure 5—figure supplement 2***). We report Pearson's correlations calculated in R.

We downloaded differential protein-level genes, effect sizes, and directional selection classifications from ***Supplementary file 4*** of ***Khan et al., 2013***. Khan et al. modeled differential protein expression of 3390 genes using high-resolution mass spectrometry of stable isotope labeling by amino acids in cell culture (SILAC) collected from five human, five chimpanzee, and five rhesus macaque LCLs.

Supplemental functional data

We downloaded human protein length (in number of amino acids) for proteins annotated as reviewed for high confidence from UniProtKB (***UniProt Consortium, 2019***). We downloaded ubiquitination protein modification data from PhoshoSitePlus version 050320 (***Hornbeck et al., 2015***). For all analyses in which we used interaction or ubiquitination data, we normalized the values by number of amino acids. To identify 3′ UTR AREs in human RefSeq annotated 3′ UTRs, we used the transcriptome_properties.py script published in ***Floor and Doudna, 2016***, available at https://github.com/stephenfloor/tripseq-analysis (***Mittleman, 2021b***; copy archived at swh:1:rev:3e823abcca5b8c1e5e89dd9bd4c49e8673b3e957) with the –au-elements flag (***Floor, 2017***; ***Floor and Doudna, 2016***). According to ***Floor and Doudna, 2016***, the fraction of AU elements is the percentage of the 3′ UTR with repeating AU elements of 5nt or more (***Floor and Doudna, 2016***).

Gene set enrichments

We performed Fast Gene Set Enrichment Analysis (FGSEA) in R (minSize = 15, maxSize = 500, nperm = 100,000) to identify enriched gene ontology (GO) terms enriched within the genes differentially expressed with at least one PAS differentially used between species. We downloaded the C5: GO terms from MSigDB. Significant GO terms can be found in ***Supplementary file 3***. To identify enriched GO terms in the remaining analyses, we used the GOrilla setting for two unranked lists of genes. To test for GO terms associated with species-specific PAS, we input the genes with a species-specific PAS as the target list and all genes with at least one identified PAS as the background set. For the differentially translated and differential protein-level background sets, we tested genes with at least one differentially used PAS as the target. We identified no significant GO terms in tdifferential protein-level analysis. Finally, among genes with at least one differentially used PAS, we tested for enriched GO terms for the genes differentially expressed in protein but not mRNA. Significant terms for each set can be found in ***Supplementary file 1***.

Acknowledgements

We thank N Gonzales for comments on the manuscript. We thank Y Li, M Ward, and G Housman for useful discussion. Funding: This work was supported by the US National Institutes of Health (R01HG010772 and R35GM13172 to YG). BEM supported by T32 GM09197 to the University of Chicago and F31HL149259 to BEM 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.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	T32GM09197	Briana E Mittleman
National Institutes of Health	F31HL149259	Briana E Mittleman
National Institutes of Health	R01HG010772	Yoav Gilad
National Institutes of Health	R35GM13172	Yoav Gilad
National Center for Advancing Translational Sciences	K12HL119995	Sebastian Pott

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Briana E Mittleman, Conceptualization, Data curation, Formal analysis, Funding acquisition, Visualization, Methodology, Writing - original draft, Writing - review and editing; Sebastian Pott, Data curation, Investigation, Methodology, Writing - review and editing; Shane Warland, Data curation, Methodology, Writing - review and editing; Kenneth Barr, Formal analysis, Methodology, Writing - review and editing; Claudia Cuevas, Resources, Data curation; Yoav Gilad, Conceptualization, Resources, Supervision, Funding acquisition, Methodology, Project administration, Writing - review and editing

Author ORCIDs

Briana E Mittleman https://orcid.org/0000-0002-4979-4652
Sebastian Pott https://orcid.org/0000-0002-4118-6150
Kenneth Barr http://orcid.org/0000-0002-0769-7053
Yoav Gilad https://orcid.org/0000-0001-8284-8926

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.62548.sa1
Author response https://doi.org/10.7554/eLife.62548.sa2

Additional files

Supplementary files

- Supplementary file 1. Gene ontology (GO) enrichment results. Regulatory processes, components, and functions identified through GOrilla as enriched at 5% FDR. Table includes the species-specific PAS genes (species-specific PAS), genes differentially translated with at least one differentially used PAS (TranslationandAPA), and the genes differentially expressed in protein but not mRNA that also have at least one differentially used PAS (APA and protein, no expression difference). GO term, description, p-value, and FDR q-value from GOrilla.

- Supplementary file 2. Polyadenylation sites (PAS) differential usage results. Column names as described – PAS: polyadenylation site, gene: gene (cluster in leafcutter), PAS_logeffectsize: log effect size for differential usage of the PAS between human and chimpanzee, PAS_deltaPAU: difference in polyadenylation site usage between human and chimpanzee (delta PSI in leafcutter), Gene_-logLR: log likelihood ratio for differential usage of any PAS in the gene (cluster likelihood ratio in leafcutter), Gene_adjustedP-value: adjusted p-value for differential usage of the any PAS in the gene (cluster p-value in leafcutter).

- Supplementary file 3. Fast Gene Set Enrichment Analysis (FGSEA) results for enrichment genes differentially expressed with at least one differentially used polyadenylation sites (PAS). FGSEA results using the C5: ontology gene set from the MSigDB collection. Pathway from C5 set, p-value adjusted for multiple testing, normalized enrichment score from FGSEA, and number of genes in the pathway.

- Supplementary file 4. 3' Sequencing metadata. Column names as described – Species: Cell line species, Lines: Cell line ID, Fraction: Cellular fraction, CollectionDate: Date of cell harvest and nuclear isolation, Extraction_date: Date of RNA extraction, Collection_person: Author initial for who processed cell harvest and nuclear isolation, UndilutedAverage: Average of two cell count measurements 1×10^6, AverageAlive: Average of two cell live dead counts – calculated with trypan blue stain, Concentration: Extracted RNA concentration (ng/mL), RIN: RIN score for extracted RNA, 260.280.Ratio: 260/280 ratio calculated on nanodrop, Library: 3' Seq library date, Reads: Number of sequenced reads, Mapped_wMP: Number of Mapped reads before removing reads likely due to mispriming, Mapped_Clean: Number of Mapped reads after removing reads likely due to mispriming.

- Supplementary file 5. Metadata for RNA-sequencing data. Column names as described – Species: Cell line species, Lines: Cell line ID, Collection_person: Author initial for who processed cell harvest and nuclear isolation, UndilutedAverage: Average of two cell count measurements 1×10^6, AverageAlive: Average of two cell live dead counts – calculated with trypan blue stain, CollectionDate:

Date of cell harvest and nuclear isolation, Extraction: Date of RNA extraction, RIN: RIN score for extracted RNA, BioAConc: RNA concentration (ng/µL), Reads: Number of Sequenced reads, Mapped: Number of mapped reads, AssignedOrtho: Number of mapped reads assigned to orthologous exons.

- Supplementary file 6. Differential expression results. Column names as described – Differential expression results from limma. gene: tested gene, logFC: log twofold change in normalized gene expression, adj.P.Val: BH adjusted p-value from t test, B: Beta value, t: t statistic.
- Transparent reporting form

Data availability

Sequencing data available on GEO under accession GSE155245.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Mittleman BE, Pott S, Warland S, Barr K, Cuevas C, Gilad Y	2021	Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155245	NCBI Gene Expression Omnibus, GSE155245

References

Ara T, Lopez F, Ritchie W, Benech P, Gautheret D. 2006. Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics* **7**:189. DOI: https://doi.org/10.1186/1471-2164-7-189, PMID: 16872498

Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y. 2014. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLOS Genetics* **10**:e1004663. DOI: https://doi.org/10.1371/journal.pgen.1004663, PMID: 25233095

Barreau C, Paillard L, Osborne HB. 2005. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Research* **33**:7138–7150. DOI: https://doi.org/10.1093/nar/gki1012, PMID: 16391004

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2015. Genomic variation impact of regulatory variation from RNA to protein. *Science* **347**:664–667. DOI: https://doi.org/10.1126/science.1260793, PMID: 25657249

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Research* **10**:1001–1010. DOI: https://doi.org/10.1101/gr.10.7.1001, PMID: 10899149

Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, Eres I, Cuevas C, Marques-Bonet T, Gilad Y. 2020. A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Research* **30**:250–262. DOI: https://doi.org/10.1101/gr.254904.119, PMID: 31953346

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**:708–715. DOI: https://doi.org/10.1101/gr.1933104, PMID: 15060014

Blekhman R, Oshlack A, Gilad Y. 2009. Segmental duplications contribute to gene expression differences between humans and chimpanzees. *Genetics* **182**:627–630. DOI: https://doi.org/10.1534/genetics.108.099960, PMID: 19332884

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Research* **20**:180–189. DOI: https://doi.org/10.1101/gr.099226.109, PMID: 20009012

Buccitelli C, Selbach M. 2020. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* **21**:630–644. DOI: https://doi.org/10.1038/s41576-020-0258-4

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* **187**:1225–1234. DOI: https://doi.org/10.1534/genetics.110.126177, PMID: 21321133

Calarco JA, Xing Y, Cáceres M, Calarco JP, Xiao X, Pan Q, Lee C, Preuss TM, Blencowe BJ. 2007. Global analysis of alternative splicing differences between humans and chimpanzees. *Genes & Development* **21**:2963–2975. DOI: https://doi.org/10.1101/gad.1606907, PMID: 17978102

Caliskan M, Cusanovich DA, Ober C, Gilad Y. 2011. The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics* **20**:1643–1652. DOI: https://doi.org/10.1093/hmg/ddr041, PMID: 21289059

Çalışkan M, Pritchard JK, Ober C, Gilad Y. 2014. The effect of freeze-thaw cycles on gene expression levels in lymphoblastoid cell lines. *PLOS ONE* **9**:e107166. DOI: https://doi.org/10.1371/journal.pone.0107166, PMID: 25192014

Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. 2016. Defining the consequences of genetic variation on a proteome-wide scale. *Nature* **534**:500–505. DOI: https://doi.org/10.1038/nature18270, PMID: 27309819

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69–87. DOI: https://doi.org/10.1038/nature04072, PMID: 16136131

Colgan DF, Manley JL. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes & Development* **11**: 2755–2766. DOI: https://doi.org/10.1101/gad.11.21.2755, PMID: 9353246

Csárdi G, Franks A, Choi DS, Airoldi EM, Drummond DA. 2015. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLOS Genetics* **11**:e1005206. DOI: https://doi.org/10.1371/journal.pgen.1005206, PMID: 25950722

de Bie P, Ciechanover A. 2011. Ubiquitination of E3 ligases: self-regulation of the ubiquitin system via Proteolytic and non-proteolytic mechanisms. *Cell Death & Differentiation* **18**:1393–1402. DOI: https://doi.org/10.1038/cdd.2011.16, PMID: 21372847

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Molecular Cell* **43**:853–866. DOI: https://doi.org/10.1016/j.molcel.2011.08.017, PMID: 21925375

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI: https://doi.org/10.1093/bioinformatics/bts635, PMID: 23104886

Dubbury SJ, Boutz PL, Sharp PA. 2018. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature* **564**:141–145. DOI: https://doi.org/10.1038/s41586-018-0758-y, PMID: 30487607

Dubnikov T, Ben-Gedalya T, Cohen E. 2017. Protein quality control in health and disease. *Cold Spring Harbor Perspectives in Biology* **9**:a023523. DOI: https://doi.org/10.1101/cshperspect.a023523, PMID: 27864315

Enard W, Fassbender A, Model F, Adorján P, Pääbo S, Olek A. 2004. Differences in DNA methylation patterns between humans and chimpanzees. *Current Biology* **14**:R148–R149. DOI: https://doi.org/10.1016/j.cub.2004.01.042

Eres IE, Luo K, Hsiao CJ, Blake LE, Gilad Y. 2019. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genetics* **15**:e1008278. DOI: https://doi.org/10.1371/journal.pgen.1008278, PMID: 31323043

Floor SN. 2017. tripseq-analysis. *Github*. 3e823ab. https://github.com/stephenfloor/tripseq-analysis

Floor SN, Doudna JA. 2016. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5**:e10921. DOI: https://doi.org/10.7554/eLife.10921, PMID: 26735365

Gokhman D, Nissim-Rafinia M, Agranat-Tamir L, Housman G, García-Pérez R, Lizano E, Cheronet O, Mallick S, Nieves-Colón MA, Li H, Alpaslan-Roodenberg S, Novak M, Gu H, Osinski JM, Ferrando-Bernal M, Gelabert P, Lipende I, Mjungu D, Kondova I, Bontrop R, et al. 2020. Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nature Communications* **11**:1189. DOI: https://doi.org/10.1038/s41467-020-15020-6, PMID: 32132541

Hilgers V, Perry MW, Hendrix D, Stark A, Levine M, Haley B. 2011. Neural-specific elongation of 3′ UTRs during *Drosophila* development. *PNAS* **108**:15864–15869. DOI: https://doi.org/10.1073/pnas.1112672108, PMID: 21896737

Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* **43**:D512–D520. DOI: https://doi.org/10.1093/nar/gku1267

Housman G, Havill LM, Quillen EE, Comuzzie AG, Stone AC. 2019. Assessment of DNA methylation patterns in the bone and cartilage of a nonhuman primate model of osteoarthritis. *Cartilage* **10**:335–345. DOI: https://doi.org/10.1177/1947603518759173, PMID: 29457464

Housman G, Gilad Y. 2020. Prime time for primate functional genomics. *Current Opinion in Genetics & Development* **62**:1–7. DOI: https://doi.org/10.1016/j.gde.2020.04.007, PMID: 32544775

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**:1299–1320. DOI: https://doi.org/10.1038/nature04226, PMID: 16255080

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *PNAS* **106**:7028–7033. DOI: https://doi.org/10.1073/pnas.0900028106, PMID: 19372383

Kedersha N, Anderson P. 2002. Stress granules: sites of mRNA triage that regulate mRNA stability and translatability. *Biochemical Society Transactions* **30**:963–969. DOI: https://doi.org/10.1042/bst0300963, PMID: 12440955

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Research* **12**:656–664. DOI: https://doi.org/10.1101/gr.229202, PMID: 11932250

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12**:996–1006. DOI: https://doi.org/10.1101/gr.229102, PMID: 12045153

Khaitovich P, Enard W, Lachmann M, Pääbo S. 2006. Evolution of primate gene expression. *Nature Reviews Genetics* **7**:693–702. DOI: https://doi.org/10.1038/nrg1940, PMID: 16921347

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342**:1100–1104. DOI: https://doi.org/10.1126/science.1242379, PMID: 24136357

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**:107–116. DOI: https://doi.org/10.1126/science.1090005, PMID: 1090005

Lee SH, Singh I, Tisdale S, Abdel-Wahab O, Leslie CS, Mayr C. 2018. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561**:127–131. DOI: https://doi.org/10.1038/s41586-018-0465-8, PMID: 30150773

Li Y, Sun Y, Fu Y, Li M, Huang G, Zhang C, Liang J, Huang S, Shen G, Yuan S, Chen L, Chen S, Xu A. 2012. Dynamic landscape of tandem 3' UTRs during zebrafish development. *Genome Research* **22**:1899–1906. DOI: https://doi.org/10.1101/gr.128488.111, PMID: 22955139

Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK. 2018. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics* **50**:151–158. DOI: https://doi.org/10.1038/s41588-017-0004-9, PMID: 29229983

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**:923–930. DOI: https://doi.org/10.1093/bioinformatics/btt656, PMID: 24227677

Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, John B. 2012. An in-depth map of polyadenylation sites in Cancer. *Nucleic Acids Research* **40**:8460–8471. DOI: https://doi.org/10.1093/nar/gks637, PMID: 22753024

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482. DOI: https://doi.org/10.1038/nature10530

Mayer A, Churchman LS. 2016. Genome-wide profiling of RNA polymerase transcription at Nucleotide resolution in human cells with native elongating transcript sequencing. *Nature Protocols* **11**:813–833. DOI: https://doi.org/10.1038/nprot.2016.047, PMID: 27010758

Mayr C. 2016. Evolution and biological roles of alternative 3'UTRs. *Trends in Cell Biology* **26**:227–237. DOI: https://doi.org/10.1016/j.tcb.2015.10.012, PMID: 26597575

Mayr C. 2017. Regulation by 3'-Untranslated regions. *Annual Review of Genetics* **51**:171–194. DOI: https://doi.org/10.1146/annurev-genet-120116-024704, PMID: 28853924

McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK. 2013. Identification of genetic variants that affect histone modifications in human cells. *Science* **342**:747–749. DOI: https://doi.org/10.1126/science.1242429, PMID: 24136359

Mittleman BE, Pott S, Warland S, Zeng T, Mu Z, Kaur M, Gilad Y, Li Y. 2020. Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**:e57492. DOI: https://doi.org/10.7554/eLife.57492, PMID: 32584258

Mittleman B. 2021a. Comparative APA Analysis. *Github*. Afc57f4. https://github.com/brimittleman/Comparative_APA

Mittleman B. 2021b. tripseq-analysis. *Software Heritage*. swh:1:rev:3e823abcca5b8c1e5e89dd9bd4c49e8673b3e957. https://archive.softwareheritage.org/swh:1:dir:d7c631f71dd7ab3a9d40cbce627c5fda9281e4e7;origin=https://github.com/stephenfloor/tripseq-analysis;visit=swh:1:snp:c7a7c70e5b66e638bde1708a5782ffd2d0417b34;anchor=swh:1:rev:3e823abcca5b8c1e5e89dd9bd4c49e8673b3e957/

Moll P, Ante M, Seitz A, Reda T. 2014. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nature Methods* **11**:972. DOI: https://doi.org/10.1038/nmeth.f.376

Moore AE, Chenette DM, Larkin LC, Schneider RJ. 2014. Physiological networks and disease functions of RNA-binding protein AUF1: rna-binding protein AUF1. *Wiley Interdisciplinary Reviews RNA* **5**:549–564. DOI: https://doi.org/10.1002/wrna.1230

Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, Meiners T, Müller C, Obermaier E, Prati D, Socher SA, Sonnemann I, Wäschke N, Wubet T, Wurst S, Rillig MC. 2014. Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecology and Evolution* **4**:3514–3524. DOI: https://doi.org/10.1002/ece3.1155, PMID: 25478144

Müller-McNicoll M, Rossbach O, Hui J, Medenbach J. 2019. Auto-regulatory feedback by RNA-binding proteins. *Journal of Molecular Cell Biology* **11**:930–939. DOI: https://doi.org/10.1093/jmcb/mjz043, PMID: 31152582

Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. 2011. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLOS Genetics* **7**:e1001316. DOI: https://doi.org/10.1371/journal.pgen.1001316, PMID: 21383968

Pai AA, Baharian G, Pagé Sabourin A, Brinkworth JF, Nédélec Y, Foley JW, Grenier JC, Siddle KJ, Dumaine A, Yotova V, Johnson ZP, Lanford RE, Burge CB, Barreiro LB. 2016. Widespread shortening of 3' Untranslated regions and increased exon inclusion are evolutionarily conserved features of innate immune responses to infection. *PLOS Genetics* **12**:e1006338. DOI: https://doi.org/10.1371/journal.pgen.1006338, PMID: 27690314

Pan Z, Zhang H, Hague LK, Lee JY, Lutz CS, Tian B. 2006. An intronic polyadenylation site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism. *Gene* **366**:325–334. DOI: https://doi.org/10.1016/j.gene.2005.09.024, PMID: 16316725

Patraquim P, Warnefors M, Alonso CR. 2011. Evolution of hox post-transcriptional regulation by alternative polyadenylation and microRNA modulation within 12 *Drosophila* genomes. *Molecular Biology and Evolution* **28**:2453–2460. DOI: https://doi.org/10.1093/molbev/msr073, PMID: 21436120

Pavlovic BJ, Blake LE, Roux J, Chavarria C, Gilad Y. 2018. A comparative assessment of human and chimpanzee iPSC-derived cardiomyocytes with primary heart tissues. *Scientific Reports* **8**:15312. DOI: https://doi.org/10.1038/s41598-018-33478-9, PMID: 30333510

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**:110–121 . DOI: https://doi.org/10.1101/gr.097857.109

Pruitt KD, Tatusova T, Maglott DR. 2004. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**:D501–D504 . DOI: https://doi.org/10.1093/nar/gki025

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033, PMID: 20110278

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**: W160–W165. DOI: https://doi.org/10.1093/nar/gkw257, PMID: 27079975

Ravid T, Hochstrasser M. 2008. Diversity of degradation signals in the ubiquitin-proteasome system. *Nature Reviews Molecular Cell Biology* **9**:679–689. DOI: https://doi.org/10.1038/nrm2468, PMID: 18698327

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**:e47. DOI: https://doi.org/10.1093/nar/gkv007, PMID: 25605792

Romero IG, Gopalakrishnan S, Gilad Y. 2018. Widespread conservation of chromatin accessibility patterns and transcription factor binding in human and chimpanzee induced pluripotent stem cells. *bioRxiv*. DOI: https://doi.org/10.1101/466631

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320**:1643–1647. DOI: https://doi.org/10.1126/science.1155390, PMID: 18566288

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**:849–864. DOI: https://doi.org/10.1101/gr.213611.116, PMID: 28396521

Sheppard S, Lawson ND, Zhu LJ. 2013. Accurate identification of polyadenylation sites from 3′ end deep sequencing using a naive bayes classifier. *Bioinformatics* **29**:2564–2571. DOI: https://doi.org/10.1093/bioinformatics/btt446, PMID: 23962617

Siegel DA, Tonqueze OL, Biton A, Zaitlen N, Erle DJ. 2020. Massively parallel analysis of human 3′ UTRs Reveals that AU-Rich Element Length and Registration Predict mRNA Destabilization. *bioRxiv*. DOI: https://doi.org/10.1101/2020.02.12.945063

Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, Tong L. 2018. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *PNAS* **115**:E1419–E1428. DOI: https://doi.org/10.1073/pnas.1718723115, PMID: 29208711

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research* **33**:201–212. DOI: https://doi.org/10.1093/nar/gki158, PMID: 15647503

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology* **18**:18–30. DOI: https://doi.org/10.1038/nrm.2016.116, PMID: 27677860

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**:D506–D515. DOI: https://doi.org/10.1093/nar/gky1049, PMID: 30395287

Vasudevan S, Peltz SW, Wilusz CJ. 2002. Non-stop decay–a new mRNA surveillance pathway. *BioEssays* **24**:785–788. DOI: https://doi.org/10.1002/bies.10153, PMID: 12210514

Verheijen J, Wong SY, Rowe JH, Raymond K, Stoddard J, Delmonte OM, Bosticardo M, Dobbs K, Niemela J, Calzoni E, Pai S-Y, Choi U, Yamazaki Y, Comeau AM, Janssen E, Henderson L, Hazen M, Berry G, Rosenzweig SD, Aldhekri HH, et al. 2020. Defining a new immune deficiency syndrome: man2b2-cdg. *Journal of Allergy and Clinical Immunology* **145**:1008–1011. DOI: https://doi.org/10.1016/j.jaci.2019.11.016

Wang R, Zheng D, Yehia G, Tian B. 2018a. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research* **28**:1427–1441. DOI: https://doi.org/10.1101/gr.237826.118

Wang SH, Hsiao CJ, Khan Z, Pritchard JK. 2018b. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biology* **19**:83. DOI: https://doi.org/10.1186/s13059-018-1451-z, PMID: 29950183

Ward MC, Gilad Y. 2019. A generally conserved response to hypoxia in iPSC-derived cardiomyocytes from humans and chimpanzees. *eLife* **8**:e42374. DOI: https://doi.org/10.7554/eLife.42374, PMID: 30958265

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics* **39**: 457–466. DOI: https://doi.org/10.1038/ng1990, PMID: 17334365

Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, Sun BB, Laser A, Maranville JC, Wu H, Ho JE, Courchesne P, Lyass A, Larson MG, Gieger C, Graumann J, Johnson AD, Danesh J, Runz H, Hwang SJ, et al. 2018. Author correction: genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications* **9**:3853. DOI: https://doi.org/10.1038/s41467-018-06231-z, PMID: 30228274

Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, Stephens M, Pritchard JK, Gilad Y. 2014. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biology* **15**:547. DOI: https://doi.org/10.1186/s13059-014-0547-3, PMID: 25468404