CPA-Perturb-seq: Multiplexed single-cell characterization of alternative polyadenylation regulators

Madeline H. Kowalski[1,2,3*], Hans-Hermann Wessels[1,2*#], Johannes Linder[4,5*], Saket Choudhary[1,2], Austin Hartman[1], Yuhan Hao[1,2], Isabella Mascio[1,2], Carol Dalgarno[1], Anshul Kundaje[4,5], Rahul Satija[1,2,3,#]

1 New York Genome Center, New York, NY, USA.
2 Center for Genomics and Systems Biology, New York University, New York, NY, USA.
3 New York University Grossman School of Medicine, New York, NY, USA.
4 Department of Genetics, Stanford University, Stanford USA.
5 Department of Computer Science, Stanford University, Stanford USA.

* These authors contributed equally
# Correspondence: hwessels@nygenome.org, rsatija@nygenome.org

ABSTRACT

Most mammalian genes have multiple polyA sites, representing a substantial source of transcript diversity that is governed by the cleavage and polyadenylation (CPA) regulatory machinery. To better understand how these proteins govern polyA site choice we introduce CPA-Perturb-seq, a multiplexed perturbation screen dataset of 42 known CPA regulators with a 3' scRNA-seq readout that enables transcriptome-wide inference of polyA site usage. We develop a statistical framework to specifically identify perturbation-dependent changes in intronic and tandem polyadenylation, and discover modules of co-regulated polyA sites exhibiting distinct functional properties. By training a multi-task deep neural network (APARENT-Perturb) on our dataset, we delineate a *cis*-regulatory code that predicts responsiveness to perturbation and reveals interactions between distinct regulatory complexes. Finally, we leverage our framework to re-analyze published scRNA-seq datasets, identifying new regulators that affect the relative abundance of alternatively polyadenylated transcripts, and characterizing extensive cellular heterogeneity in 3' UTR length amongst antibody-producing cells. Our work highlights the potential for multiplexed single-cell perturbation screens to further our understanding of post-transcriptional regulation *in vitro* and *in vivo*.

INTRODUCTION

RNA cleavage and polyadenylation represent post-transcriptional regulatory mechanisms that are required for the maturation of eukaryotic pre-mRNA[1–4]. The majority of mammalian genes harbor multiple polyA sites, enabling a single gene to encode multiple mRNA transcripts via alternative polyadenylation[5–7]. The distinct 3' ends arising from this process add to the rich diversity of mammalian transcriptomes, and can influence multiple distinct stages of the RNA life cycle. For example, shortening of the 3' untranslated region (UTR) at tandem UTRs can affect transcript stability and localization[8,9], while alternative polyadenylation at intronic sites can lead to the generation of truncated coding or non-coding transcripts[10–12]. More generally, widespread changes in polyadenylation have been demonstrated in many biological contexts including cellular proliferation[13], tumorigenesis[14,15], embryonic development[16], and secretory cell differentiation[17].

Biochemical and molecular studies have revealed a subset of core and accessory proteins that are responsible for regulating polyA site choice. For instance, the cleavage and polyadenylation specificity factor complex (CPSF) catalyzes cleavage, the Cleavage factor I (CFIm) and Cleavage factor II (CFIIm)

complexes bind auxiliary recognition sequences, and PolyA Polymerase (PAP) is responsible for adding the polyA tail[4]. While the identity of key proteins is known, our understanding of how their relative concentration and interaction with RNA sequence elements influences alternative polyadenylation remains incomplete, representing a key challenge for our understanding of post-transcriptional regulation.

Functional genomics approaches offer exciting potential to address these questions. Recently, massively parallel reporting assays (MPRA) combined with deep neural networks have been utilized to construct sequence-based models of alternative polyadenylation and can successfully predict cleavage site usage under baseline conditions[18–21]. Alternatively, genome-wide 3' transcriptome technologies can be used to profile changes in polyA site usage across different biological samples[16,22–24], including those that perform genetic perturbations of CPA regulators. While some individual studies perturb individual or small sets of regulators[25–28], others have used siRNA-based screening approaches to generate larger resources[29,30]. Alternatively, multiplexed single-cell technologies like Perturb-seq leverage single-cell RNA-sequencing (scRNA-seq) for high-throughput transcriptome-wide characterization of molecular perturbations[31–34]. While scRNA-seq is typically applied to profile heterogeneity in gene expression levels, these data can also be leveraged to characterize changes in transcript structure. In particular, the majority of scRNA-seq protocols are explicitly designed to capture the 3' end of polyadenylated mRNA transcripts. Therefore, these methods are well-suited to quantify transcriptome-wide polyA site usage at single-cell resolution alongside gene abundances, revealing dynamic changes in polyadenylation during cellular differentiation and disease[35–38].

Here, we introduced CPA-Perturb-seq, a resource where we perturb known regulators of CPA in a multiplexed 3' scRNA-seq screen, and quantify each perturbation's effect on polyA site usage at single-cell resolution. We introduce new computational tools to specifically quantify changes in polyA site usage in sparse single-cell datasets, and to decouple these from changes in gene abundance levels. Our analyses reveal substantial diversity in the number and types of polyA sites affected by perturbation of different regulators, modules of sites that are co-regulated across perturbations, and the role of interacting RNA sequence elements in determining polyA site selection. We also demonstrate how our computational tools can be applied to any 3' scRNA-seq dataset, identify new regulators in a genome-scale Perturb-seq resource[39], and characterize natural variation in alternative polyadenylation amongst high-resolution subsets of antibody-secreting plasma cells. Together, our analyses demonstrate how single-cell sequencing can move beyond gene expression analyses and improve our understanding of post-transcriptional gene regulation.

RESULTS

Multiplexed Perturb-seq screens of 3' polyA site usage

We sought to understand how systematic perturbations of genes involved in cleavage and polyadenylation would affect alternative polyadenylation at single-cell resolution (Figure 1A). We designed a library of 162 single guide RNAs (sgRNAs) targeting 42 genes and 10 non-targeting (NT) controls (Supplementary Table 1). Our target set included 18 genes that are known members of core cleavage and polyadenylation complexes, including the Cleavage Factor Im (CFIm), Cleavage Factor IIm (CFIIm), Cleavage and polyadenylation specificity factor (CPSF), and and Cleavage stimulation factor (CSTF) complexes (Figure 1B). We also included 23 genes that have been previously implicated in affecting relative polyA site usage, including subunits of the PAF complex[40], the splicing factor

SRSF3[27], and THOC5[41], a member of the transcription/export complex (TREX) (Supplementary Table 1).

We performed a pooled CRISPR inhibition screen (Supplementary Methods) in HEK293FT cells and used the Perturb-seq experimental workflow to simultaneously capture the identity of the guide each cell received along with a 3' scRNA-seq readout (Figure 1A). While we focus our primary analyses on the deeply profiled HEK293FT dataset (median of 1,788 cells per perturbation), we also repeated the experiment in K562 cells to obtain a second biological context (median of 596 cells per perturbation). Across two biological replicates (independent viral transductions) in each cell line, we obtained a total of 109,661 single cells (Supplementary Table 2) where we were able to successfully assign a single gRNA.

We utilized the scRNA-seq data to quantify both gene expression and transcriptome-wide polyA site usage profiles for each cell. We used tools from the polyApipe pipeline to first identify a set of possible cleavage and polyA sites, and then to quantify their usage in single cells[38]. We further restricted our analysis to polyA sites that are within 50 nucleotides of polyA sites identified in polyAdbv3[7], a database of polyA sites generated from multiple human cell lines (Supplementary Methods). We only included sites located within an intron or the last exon of a gene (Supplementary Methods). This strategy focuses specifically on splicing-independent changes in alternative polyadenylation, and does not consider changes in alternative last exon usage driven by splicing.

We identified a total of 33,399 polyA sites across 12,194 detected genes, and found that 8,077 genes exhibited usage of two or more polyA sites in our dataset (5,253 genes exhibited usage of three or more) (Supplementary Figure 1A). Moreover, we found that 76% and 79% of our identified exonic and intronic polyA sites, respectively, contained the canonical AATAAA/ATTAAA cleavage motif in the region 50 bp upstream of the inferred cleavage site, as would be expected for *bona fide* polyA sites (Supplementary Figure 1B). We therefore assigned reads from our 3' scRNA-seq dataset to each polyA site (Supplementary Methods), generating a polyA site/cell count matrix for downstream analysis.

Multiple groups have previously observed that datasets from pooled single-cell CRISPR screens often contain confounding sources of variation[32,33,42]. These include heterogeneity across replicate experiments, cell cycle differences, or variable perturbation efficiency even amongst cells expressing the same gRNA. We applied our previously developed computational pipeline, Mixscape[42], to address these effects and to remove cells that exhibit no phenotypic evidence of perturbation (Supplementary Methods). For 16 of 42 regulators, Mixscape classified all cells as 'non-perturbed', suggesting that even if the perturbation was successful (Supplementary Figure 1C), the global effect on the transcriptome was minimal. For the remaining 26 genes, Mixscape classified 76% of cells as perturbed.

Perturbed cells exhibited diverse changes in alternative polyadenylation. For example, at the CBX3 locus, perturbation of NUDT21 and CPSF6 shifted expression towards the proximal polyA site ('3' UTR shortening') RBBP6 and PCF11 perturbation shifted expression towards the distal isoform ('3' UTR lengthening'), while the regulators FIP1L1 and CPSF3L did not induce changes (Figure 1C). These changes were reproducible across biological replicates and multiple independent gRNA (3-4 per gene; Supplementary Figure 1D). We used the polyA site/cell count matrix for perturbed cells along with 3,789 NT controls, as input to linear discriminant analysis (LDA), UMAP visualization (Figure 1D), and unsupervised clustering of the polyA site matrix (Supplementary Figure 1E). These analyses revealed that cells clustered not only by the perturbation they received, but also into broader complexes. For

example, cellular profiles after NUDT21 and CPFS6 (both members of the CFIm complex) perturbation were highly correlated, as were profiles for members of the CPSF (CPSF1-4, FIP1L1), CSTF (CSTF1/3), and PAF (PAF1, CTR9, LEO1, CDC73) complexes.

These results suggest our dataset can be used to uncover complex-specific 'modules' of co-regulated polyA sites, each of which are responsive to perturbation by a set of functionally related regulators. However, we note that changes in the polyA site/cell count matrix can reflect both changes in 3' UTR utilization, but also changes in the overall abundance of the gene even in the absence of isoform-level changes. We observed both cases in our dataset. For example, when perturbing CSTF3 (Figure 2A-D), we identify cases where changes in the utilization of a gene's proximal peak corresponds exclusively to a change in total RNA abundance (ATP6V1G1), exclusively to a change in transcript length due to 3' UTR shortening (HNRNPH3), or changes in both abundance and relative isoform usage (MRPS16).

Quantifying relative polyadenylation levels at single-cell resolution

To specifically characterize perturbation-driven effects on alternative polyadenylation, we therefore sought to design a computational approach to deconvolve these two effects. While computing ratios of polyA counts for each site within a gene is typically used to study alternative polyadenylation in bulk analyses, computing these ratios in scRNA-seq data is typically infeasible or noisy due to data sparsity. Instead, for each polyA site in each single cell, we aimed to model and quantify the degree of over or under-utilization, compared to the expected usage observed in NT cells.

We note that this problem is conceptually similar to quantifying the degree of increased or decreased expression of each gene in each cell, compared to the population average. We and others have developed statistical methods to address this challenge for gene abundances in scRNA-seq data using generalized linear models[43–46]. Here, we chose to extend this framework to model alternative polyadenylation (Figure 2E). We utilized the Dirichlet multinomial distribution to model the background distribution of each polyA site in NT control cells. The expected value for each site is set by the relative usage of all polyA sites within a gene, which controls for the overall expression of the gene, and can be robustly estimated from 3,789 NT cells.

When quantifying the variance for each site, the Dirichlet multinomial allows for the possibility of overdispersion compared to the standard multinomial[47], analogous to the routine use of the negative binomial distribution to model Poisson overdispersion when modeling gene abundances[48,49]. This overdispersion accounts for natural biological heterogeneity and 'intrinsic' noise that occurs within the background population[46,50], and can be estimated directly from each dataset. As in sctransform[45], we first parameterize overdispersion estimates individually for each polyA site, but then regularize these estimates across similar sites (Supplementary Methods). The output of our procedure is a statistical model for each polyA site, describing its background usage across NT control cells.

Finally, we utilized these background models to quantify relative polyA site usage at single-cell resolution. By comparing the observed counts at each site in each cell with the expected value and variance from the Dirichlet multinomial model, we compute a Pearson residual ('polyA-residual') at each polyA site. The sign and magnitude of this residual describes the cell's relative deviation from the expected background distribution for each polyA site. A positive residual reflects that a polyA site is used more frequently in a single cell relative to the background distribution, and a negative residual reflects that a polyA site is used less frequently than the background distribution.

Our quantified polyA-residuals can be used as input for differential polyadenylation analysis, allowing us to identify polyA sites whose relative frequency changes across groups of cells while mitigating any confounding changes in overall gene abundance. We tested for changes in polyA-residuals using a linear model, including gRNA identity as a covariate, to identify reproducible changes for each perturbation (Supplementary Methods). When applied to our previous examples (Figure 2F-H), this approach successfully distinguished loci where we observed either changes in transcript structure, abundance, or both. The polyA-residuals can also be used as input for clustering and visualization. When repeating our LDA-based visualization procedure and clustering analyses on the polyA-residual matrix (Supplementary Figure 3A), we replicated our previously observed findings (Figure 1D) confirming that these co-regulatory patterns were driven by coordinated changes in polyadenylation.

Characterizing perturbation-dependent changes in polyadenylation

We next characterized the effect of each perturbation individually. We identified 6,734 genes that exhibited differential alternative polyadenylation (at least one polyA site with differential usage) in at least one the 26 gene perturbations, but observed substantial differences across regulators. CFIm complex members such as NUDT21 exhibited the strongest perturbation responses affecting more than 5,600 genes (Figure 3A), including 2,397 genes where we exclusively detected relative changes in at least one polyA site (40%; blue bar), 1,058 genes where we exclusively detected changes in total transcript levels (18%; light blue bar), and 2,536 genes where we detected changes in both abundance and structure (42%, green bar).

Our results demonstrated that differential analysis of our polyA-residuals represents an effective workflow to identify specific changes in relative polyA site usage, as opposed to total transcript abundance. For example, we observed that perturbation of PABPC1, affected the expression level of 1,265 genes, but had negligible effects on relative polyA site usage in either our HEK293FT or K562 cell dataset (Figure 3A; Supplementary Figure 3C). This result is consistent with the known cytoplasmic localization of PABPC1, which binds the polyA tail after nuclear export, and is not expected to regulate polyA site choice[51].

We next classified each significant change in polyA site usage as reflecting either intronic polyadenylation or tandem polyadenylation, based on site annotations in the polyADB database (Figure 3B). We found that in most cases, perturbing an individual regulator primarily led to changes in tandem polyA site usage. However, for a minority of regulators, such as polymerase-associated factor (PAF) complex members[52] or the RNA PolII elongation factor SCAF8[26], responses were primarily associated with intronic polyadenylation in both HEK293FT and K562 cells (Supplementary Figure 3D). As both sets of regulators interact with RNA Polymerase and play an established role in regulating polymerase progression, these results provide additional evidence for kinetic models where changes in elongation rate can influence alternative polyadenylation[53], and suggest that this relationship is particularly important in the context of intronic sites.

While we identified multiple regulators that primarily affected intronic polyadenylation, we found that they regulated distinct intronic sites. (Figure 3C-E). Moreover, we found that the distance between two adjacent cleavage sites in a transcript was predictive of the responsiveness to PAF1 perturbation (Supplementary Figure 3E). This relationship was strongest for polyA sites located in the first intron, but also held for downstream sites as well (Supplementary Figure 3E). However, when performing similar

analyses for SCAF8, we observed a weaker predictive power for both intronic location or distance between cleavage sites (Supplementary Figure 3F). These results demonstrate that while transcriptional elongation rate likely influences intronic polyA site selection, polymerase-interacting factors can exhibit distinct regulatory effects.

Focusing next on alternative polyadenylation between tandem polyA sites, we found that perturbation of CPA regulators resulted in striking shifts in the utility of either proximal or distal polyA sites, with most perturbations (18/26) exhibiting a skew of greater than 70% in either direction (Figure 3F). These relationships replicated in our independently obtained K562 dataset as well (Supplementary Figure 3G). We did observe a general trend where 3' UTR shortening was associated with an increase in total gene abundance (Figure 3G), consistent with the broad association between 3' UTR length and the presence of regulatory elements that may impact RNA stability[54,55].

Our observed patterns of shortening/lengthening were concordant with previous studies that utilize bulk 3' end sequencing technologies, but highlighted the advantages of the Perturb-seq technology. For example, four previous studies[25,29,30,56] have consistently revealed that perturbation of NUDT21 perturbation affects polyA site usage in a subset of genes (ranging from 375-1,600) and leads to 3' UTR shortening at tandem UTRs. In our dataset (Figure 3F; Supplementary Figure 3G), we observed more than 5,500 genes exhibiting significant changes in polyA site usage after perturbation, exhibiting not only high sensitivity but also high specificity in both HEK and K562 datasets (>93% of tandem UTR changes resulted in shortening, indicating that these reflect *bona fide* perturbation responses).

Similarly, RBBP6 perturbation has been associated with 3' UTR lengthening, but the degree of this preference (ranging from 60% to 78%) and the number of genes (ranging from 100 to 1,300) varies across studies[28–30]. In our dataset, we observed more than 2,600 genes with polyA site changes, with a high specificity (>95% lengthening at 3' UTR) in both cell lines (Figure 3F; Supplementary Figure G). These results highlight how pooled single-cell CRISPR screens, which avoid batch effects by multiplexing all perturbations and controls together, can yield accurate perturbation signatures especially when performed with high cell number and utilizing multiple independent gRNA. Moreover, this experimental design is ideally suited for the identification of co-regulated polyA sites across multiple regulators, without having to compare datasets generated across different experiments or studies.

Modules of co-regulated polyA sites exhibit distinct functional properties

While our previous analyses characterized regulators individually, we also clustered perturbations based on the observed changes to all differentially polyadenylated sites quantified by our model (Supplementary Methods; Figure 4A). Consistent with expression-based analysis, perturbation clusters reflected membership structure of core CPA complexes, as well as additional evidence of co-regulation. For instance, RBBP6, FIP1L1, and PCF11 are not members of the same complex, but all cause 3' UTR lengthening at overlapping sites upon perturbation, and cluster together. Moreover, we repeated these analyses on the K562 dataset and observed highly concordant correlation patterns (Figure 4B), suggesting these reflect co-regulatory relationships that generalize beyond a single biological context.

We found that the correlation structure was not exclusively driven by global preferences towards shortening and lengthening, but also local differences in the specific sites affected by each regulator. For example, perturbation of RBBP6 (preference towards 3' UTR lengthening) and CFIm complex members CPFS6 and NUDT21 (preference towards 3' UTR shortening) showed strongly anti-correlated

responses, reflecting their globally opposing regulatory tendencies at the same set of loci. By contrast, CSTF and CPSF complex members (preference towards 3' UTR lengthening) showed only weak anti-correlation with CFIm members, reflecting more complex patterns of co-regulation.

To further explore this, we considered a group of 1,208 genes that exhibited transcriptional shortening after CFIm perturbation (Supplementary Methods). When further subdividing this set of sites based on their response to CSTF perturbation, we observed an expected module (Figure 4C-E, Module A) of 245 polyA sites (20%) where CSTF perturbation resulted in an opposing lengthening response. However, we also identified a module (Module B) of 110 (9%) genes where CSTF perturbation also resulted in shortening, phenocopying CFIm perturbation despite their opposing global preferences. The remaining 71% of sites did not exhibit changes in utilization upon CSTF perturbation. While we identified these modules in our HEK293FT dataset, we independently observed reproducible patterns at the same loci in K562 cells (Supplementary Figure 4A).

Strikingly, we found that these gene modules exhibited clear functional differences (Figure 4F). In particular, we found that genes where we observed opposing regulatory effects between the two complexes (Module A) strongly favored the usage of proximal sites in NT cells, while genes exhibiting consistent regulatory effects (Module B) were strongly biased towards distal site usage. These results were consistent in both HEK293FT and K562 cells (Supplementary Figure 4B) and indicate that local effects, likely determined by differences in sequence content, establish the responsiveness to CSTF perturbation, and are important in establishing the proximal versus distal bias for individual genes. More broadly, we conclude that our polyA residuals represent an effective statistical approach for characterizing the perturbation responses of individual regulators, and for identifying modules of polyA sites that are co-regulated across perturbations.

APARENT-Perturb reveals an interactive cis-regulatory code

Our identification of modular patterns of differential polyadenylation that reproduce across cell types emphasizes the role of local sequence drivers in determining an individual polyA site's responsiveness to distinct perturbations. Motivated by the success of deep learning models in accurately predicting genome-wide patterns of alternative polyadenylation in baseline conditions[18–20,57–59], we sought to extend these models to predict the perturbation responses observed in our dataset. For example, APARENT2 represents a residual neural net, originally trained on MPRA datasets, that can predict baseline polyA site usage in HEK293FT cells and interpret specific sequences elements and genetic variants that drive model accuracy[57]. The ability to successfully capture nonlinear interactions, including positional and combinatorial interdependencies between motifs, highlights the ability of these models to learn intricate cis-regulatory determinants[60].

We therefore hypothesized that sequence-based learning models could predict the response of each polyA site to each of the ten highest magnitude perturbations in our dataset. To test this, we first used the pre-trained APARENT2 model to provide baseline predictions for polyA site usage. We then trained a new neural network (APARENT-Perturb) to predict usage in our Perturb-seq data, using 200nt sequences centered on the site of 3' cleavage, along with the baseline APARENT predictions as input (Figure 5A). This approach was inspired by the MTSplice model[61], and represents an ensemble-based multi-task perturbation network that can not only predict relative polyA site usage in NT control cells (baseline), but also can predict polyA site usage after perturbation. After training, APARENT-Perturb could accurately predict the isoform proportion of polyA sites for held-out genes in both the

non-targeting (NT) condition (R_S = 0.70) and in perturbations (0.65 ≤ R_S ≤ 0.73 depending on perturbation), as measured by 10-fold cross-validation (Figure 5B, Supplementary Figure 5A). When predicting relative differences in polyA site usage between a given perturbation and the NT condition, the performance varied more as some perturbations resulted in an only moderate change to APA levels (0.27 ≤ R_S ≤ 0.59), but these results were still highly significant ($2.25x10^{-125}$ ≤ p ≤ $2.66x10^{-17}$).

To interpret the model, we performed *in silico* mutagenesis (ISM) by simulating local sequence alterations and comparing the resulting predictions to the unaltered model. This procedure yields a set of nucleotide-level 'attribution scores', reflecting the contribution of each individual base to the model's prediction[62,63]. Importantly, by subtracting scores of the NT (baseline) output, we isolate each sequence's importance in predicting perturbation responses. For example, the attribution scores of the distal polyA site in the KMT5A gene, highlight an upstream UGUA motif that is predicted to drive responsiveness to NUDT21 perturbation, and a distinct downstream GU-rich region motif that drives responsiveness to CSTF3 perturbation (Figure 5C, Supplementary Figure 5B). For each perturbation, we averaged ISM scores across loci to identify regions that harbored important sequence elements (Figure 5D-E). We next used a motif discovery tool, TF-MoDISco[60,64], to cluster the attribution scores of each perturbation into a set of salient motifs (Figure 5D, Supplementary Figure 5C-D). These results recapitulate and extend previously established binding motifs and positions[2,4,59] , for example, NUDT21 and CPSF6 both display high average importance in the upstream region of polyA sites and are sensitive to UGUA motifs with T- or A-rich flanks, while CSTF1 and CSTF3 display a peak of importance in the downstream region with U- or GU-rich sequences among their top motifs.

Intriguingly, APARENT-Perturb attribution scores suggest motifs that help to coordinate joint activities of both CFIm and CSTF complex members. Our model's attribution scores predicted that CFIm perturbation responses are predicted not only by sequences upstream of the cleavage sequence, but also by a sequence element located approximately 30-50 bp downstream (downstream element; DSE). This DSE overlaps with a region of predicted importance for CSTF perturbation, reflecting a co-enrichment of functional sequences for both complexes at the same sites (Supplementary Figure 5E). While APARENT-Perturb predicted positive attribution scores for most sites in this region after NUDT21 perturbation (Supplementary Figure 5F; Decile 10), a subset (Decile 1) exhibited negative attribution scores. Indeed, we found that these two groups of sites differed in their responsiveness to CSTF1 and CSTF3 (Supplementary Figure 5G). The fact that the NUDT21 perturbation model ascribes importance to sequence motifs that drive CSTF regulation is strong evidence of sequence-driven interaction between these factors.

We had previously observed that CSTF and CFIm complex members can jointly regulate polyA sites in either the same or opposing directions (Figure 4C-F), and we identified a link between our model's attribution scores and our previously identified modules. Specifically, we found that in genes where CFIm perturbation led to transcriptional shortening and CSTF perturbation led to lengthening (Module A), the DSE at the proximal polyA site was characterized by sequence elements with high CSTF attribution scores, which are predicted to facilitate CSTF binding and regulation. However, at genes where perturbation of both complexes led to transcriptional shortening (Module B), the proximal sites exhibited significantly weaker sequence elements (Figure 5F left, p< $2.0*10^{-5}$, Wilcoxon two-sided rank sum test). By contrast, we observed increased attribution scores for Module A genes at distal sites (Figure 5F right, p< $1.6*10^{-4}$).

Taken together, these findings suggest a model where the sequence content at proximal polyA sites is particularly important both in establishment of the proximal/distal bias, as well as the responsiveness to multiple perturbations. In a subset of genes (Module A), proximal peaks contain strong motifs that serve to recruit CSTF. This regulatory structure promotes cleavage at the proximal site under baseline conditions, but leads to transcriptional lengthening after CSTF perturbation. Alternatively, a distinct gene subset (Module B) exhibits weaker sequence features at the proximal site, while the distal site contains sequences that promote recruitment of CSTF and CFIm. These loci exhibit distal cleavage under baseline conditions, but perturbation of either complex results in transcriptional shortening.

Finally, as neural networks trained on ChIP-seq data have been recently shown to successfully learn the syntax of cooperative binding between transcription factors[60], we aimed to identify similar types of interactions between CPA regulators. We used APARENT-Perturb to simulate either individual or pairwise motif insertions, and compared the predicted results to identify epistatic interactions. For example, the CFIm complex includes a NUDT21 homodimer, but it is unclear if and how multiple UGUA motifs affect binding[65,66]. We found that two adjacent UGUA motifs tended to act cooperatively in predicting the responsivity to NUDT21 perturbation. However, we only observed synergistic effects when both motifs were surrounded by GC-rich sequences, while an AT-rich context was associated with sub-additive interactions (Figure 5G, Supplementary Figure S5H-I). We also identified that two distinct sequence elements, the canonical core hexamer, and GU-rich sequence element observed in the DSE, also exhibited epistasis in predicting polyA site usage after RBBP6 perturbation. While previous work has associated that both of these motifs independently are associated with RBBP6 regulation[28], APARENT-Perturb identified a position-dependent relationship, with a maximum epistatic interaction observed when the motif distance was approximately 20bp (Figure 5H, Supplementary Figure 5J). We verified each of these results using polynomial feature regression (Supplementary Figure 5K-L). We conclude that deep learning models can be successfully applied to analyze high-throughput Perturb-seq datasets, and can reveal a cis-regulatory landscape that encodes complex patterns of co-regulation across multiple complexes.

Identification of APA regulators from genome-wide screening datasets

While the genes selected for our screen encompassed previously identified regulators of alternative polyadenylation, our computational workflow is capable of characterizing changes in polyA site usage for any 3' scRNA-seq dataset. We therefore reanalyzed a recently published genome-wide Perturb-seq dataset (GWPS)[39] which perturbed 9,866 transcriptionally active genes in K562 cells (including all 26 perturbed regulators in our datasets), but did not explore perturbation-dependent changes in polyA site usage. To address this, we computed polyA-residuals for each cell, and used these as input to differential polyadenylation analysis (Supplementary Methods). As the GWPS dataset contained far fewer cells per perturbation (median 91 cells, vs. 1,032 for in our dataset for the 26 overlapping perturbations), we identified substantially fewer genes exhibiting changes in polyA site usage (median of 165 genes per overlapping perturbation, compared to 1,351 in our data). However, even at shallow depth, the GWPS dataset enabled accurate global characterization of each regulator. For example, we observed a strong concordance in the global bias towards 3' UTR shortening or lengthening induced by regulatory perturbation across both datasets (Figure 6A).

We therefore extended our analyses to focus on a previously annotated set of 1,280 RNA binding proteins[67] in order to facilitate identification of regulators that directly modify RNA. While most perturbations exhibited minimal transcriptome-wide changes in alternative polyadenylation, we

identified 172 regulators whose perturbation affected polyA site usage in at least 50 genes (Figure 6B, Supplementary Table 4). We also identified groups of highly correlated perturbations that were consistent with and substantially expanded our previous observations (Figure 6C, Supplementary Figure 6A). For example, one group included the CFIm complex members CPSF6 and NUDT21, but also THOC3, a Transcription-Export (TREX) complex member, and TCERG1 (transcriptional elongation regulator 1). While perturbation of this module was associated with shortening at tandem 3' UTR (Supplementary Figure 6B-C), we identified a separate lengthening-associated module (Module 4; Supplementary Figure 6F) consisting of Up-frameshift complex members (UPF1, UPF2), the small ribosomal subunit (RPS24, RPS4X), and the ribosome maturation factor TSR2. While these genes are well-studied regulators of translational control and RNA stability, none have been previously associated with regulating polyA site selection. Components of the large ribosomal subunit formed were also associated with polyA site selection, but formed a separate module (Module 2; Supplementary Figure 6D), along with the translation initiation factor EIF6. These analyses demonstrate how large-scale perturbation screens can identify novel regulatory factors and suggest tight regulatory crosstalk linking changes in alternative polyadenylation with multiple processes in the RNA life cycle.

We additionally identified a third group of 15 correlated perturbations (Module 3), 13 of which have been previously identified as members of the poly(A) tail exosome targeting (PAXT) complex[68]. Perturbation of this module was associated primarily with the indirect up-regulation of intronically polyadenylated transcripts (Figure 6B), whose abundance was not changed in response to perturbation of the nuclear exosome targeting (NEXT) complex or the CPA machinery (Figure 6D). This response is driven by the PAXT complex's role in degrading prematurely terminated RNA transcripts, which accumulate in the cytoplasm after PAXT perturbation[68,69], although the nuclear surveillance machinery that specifically distinguishes premature transcripts remains unknown[70]. Intriguingly, the nuclear cap-binding complex member NCBP2 and and splicing regulator MBNL1 were also members of this module but neither are members of the PAXT complex. While NCBP2 is known to promote successful RNA export[68,71], MBNL1 perturbation has been previously linked to regulating levels of intronic retention[72], including in cases where retention leads to premature termination[73]. The striking phenocopying between perturbation of MBNL1 and PAXT subunits (Figure 6D-F) suggests a hypothesis where a splicing regulator, via its role in regulating intron retention, may assist PAXT in selecting unstable and undesired transcripts for degradation.

scRNA-seq profiles reveal extensive plasma cell heterogeneity in polyA site usage

While our previous analyses focused on cellular heterogeneity across *in-vitro* perturbation experiments, we next asked whether our statistical framework could quantify and interpret APA heterogeneity in an *in-vivo* context. For example, recent work using bulk RNA-seq datasets has demonstrated that secretory cell differentiation is associated with widespread changes in polyadenylation[17]. To further explore this, we calculated polyA-residuals on a 3' scRNA-seq dataset consisting of 49,958 circulating human peripheral blood mononuclear cells (PBMC) which includes seven COVID-19 infected samples that exhibit an induction of antibody-secreting plasma cells[74]. We processed this data to identify and quantify 20,067 polyA sites, and quantified both gene expression levels as well as polyA residuals.

Unsupervised analysis of polyA residuals revealed that heterogeneity in polyA site usage across immune subsets was relatively modest compared to heterogeneity in gene expression (Figure 6G). However, consistent with previous reports[11,17], we did observe that plasma cells exhibited clear differences compared to all cell types, including developmentally related B cell subsets. In addition to

observing a shift towards the shorter isoform of the IGHM locus (one of the first described examples of alternative polyadenylation[75]; Supplementary Figure 7A), we identified 1,8783 genes (630 intronic changes and 1253 tandem changes) exhibiting differential usage of at least one polyA site in plasma cells (Supplementary Methods). Genes exhibiting differential tandem polyadenylation were primarily associated with 3' UTR shortening (95%). Gene Ontology analysis revealed a strong enrichment for Golgi vesicle transport and protein localization, which are linked to the core secretory phenotypes of plasma cells (Figure 6H).

One key advantage of single-cell measurements is the ability to explore how multiple levels of granularity in cell annotation affect downstream analyses. We found that when further subdividing plasma cells to annotate short-lived and highly proliferative plasmablast subpopulations, these cells exhibited the most striking shifts in polyA site usage (Figure 6I-J, Supplementary Figure 7B). Surprisingly, we found that non-cycling plasma cells not only exhibited weaker changes, but also exhibited a bimodal distribution in their polyA-residuals, enabling further subdivision into two groups based on the degree of 3' UTR shortening (Figure 6J; Supplementary Figure 7C-D). These two groups differed not only in transcript structure, but also in the expression of a module of genes that were highly enriched for their involvement in respiratory and metabolic processes (Supplementary Figure 7E). These results extended previous bulk RNA-seq based findings[17], but were uniformly consistent across 11 donors (Supplementary Figure 7F). They demonstrate that widespread 3' UTR remodeling occurs in the earliest stages of plasma cell differentiation, but substantial cellular heterogeneity in polyA site usage remains even after commitment to this lineage. Future experiments will establish whether the metabolic changes observed between these groups relate to the secretory capabilities or lifespan of these cells.

We conclude that 3' scRNA-seq data can be combined with tailored computational pipelines to explore cellular heterogeneity in polyA site usage for both in vitro and primary samples, and have developed an open-source R package PASTA (PolyA Site analysis using relative Transcript Abundance) that implements the analytical methods described in this manuscript. PASTA is fully compatible with our analytical toolkit Seurat[76], and the software release includes a vignette demonstrating how users can explore changes in their datasets using PASTA and Seurat. These data and code resources will facilitate the characterization of heterogeneous alternative polyadenylation in diverse biological systems and a deeper understanding of the sequences and regulatory factors that govern post-transcriptional regulation.

DISCUSSION

In this study, we aimed to understand how the abundance of CPA regulators, as well as the presence of RNA sequence elements, affect the regulation of alternative polyadenylation across the transcriptome. We demonstrate that the Perturb-seq technology, which has been widely utilized to study transcriptional regulatory networks, can be successfully applied to study post-transcriptional regulation as well. We introduce a statistical framework to quantify changes in relative polyA site usage at single-cell resolution, and demonstrate how this approach can characterize the effect of individual regulators, identify modules of co-regulated polyA sites, and enumerate subpopulations of cells that exhibit changes in polyA site usage in any 3' scRNA-seq dataset.

Our CPA-Perturb-seq dataset revealed striking heterogeneity in the perturbation responses of different regulators. This was reflected in the number, type, and directionality of changes induced by each

perturbation. However, our dataset highlights that regulation of alternative polyadenylation is not a uniform or global process, where all polyA sites are sensitive to perturbation by all core regulators. Instead, we consistently observed evidence for modularity and substructure in our data. Even when perturbing regulators of the core CPA machinery, we identified groups of polyA sites that were co-regulated by a subset, but not all, regulators. Moreover, we identified cases where the same set of perturbations resulted in opposing responses for distinct modules of polyA sites. Our Perturb-seq dataset is well-suited for module characterization, as the multiplexed design mitigates experimental batch effects and avoids the need to compare perturbation profiles generated from different experiments or studies.

By interpreting our multi task deep neural network, APARENT-Perturb, we find that this local regulatory structure is encoded in part by sequence-specific elements that surround the cleavage site. Previous models trained on MPRA data constitute a powerful approach to identify functionally important sequence elements, but it is challenging to understand how they exert regulatory effects. By integrating these models with our perturbation data, we learn direct associations between sequence elements and regulators, providing a more mechanistic understanding of cis-regulatory element function. We demonstrate the ability of this approach to identify interactions between different regulators of alternative polyadenylation, but this approach could also be extended to deep neural networks that predict chromatin accessibility levels from DNA sequence, and to provide deeper functional interpretation of sequence variants.

While our analyses aimed to focus on regulatory mechanisms that influenced cleavage and polyadenylation decisions, we repeatedly observed cases where additional regulatory processes in the RNA life cycle would alter the relative abundance of alternatively polyadenylated transcripts. For example, we observed regulator-specific patterns that connected changes in RNA polymerase elongation rate with altered usage of intronic polyA sites. More broadly, we also found that perturbation of proteins with well-characterized roles in RNA export, RNA translation, and RNA splicing and intron retention also resulted in differential usage of polyA sites. These results highlight the extensive interdependencies that connect different RNA regulatory processes. To this end, future work may be able to exploit these interdependencies to infer RNA kinetic parameters from 3’ scRNA-seq data, for example, utilizing changes in the usage of intronic polyA sites to infer cell type-specific changes in RNA elongation rate. More broadly, our statistical method may be extended to characterize additional sources of transcriptomic diversity, such as changes in splicing from full-length datasets, in order to characterize a broader realm of post-transcriptional regulatory events.

While our study independently explores datasets deriving from either multiplexed perturbation screens or primary human samples, looking forward, we believe these contexts will be mutually informative. Functional genomics tools like Perturb-seq are especially well-suited to identify causal relationships between molecular regulators and their targets. In contrast, comparative analysis of alternative polyadenylation across biological samples, conditions, and disease states is a powerful approach for identifying transcriptome-wide changes, but identifying the causal regulators driving these responses remains challenging. We envision that the molecular signatures inferred from experiments where causal relationships are established represent important resources to interpret molecular signatures where causal relationships are unknown. These links will be particularly informative as Perturb-seq experiments extend beyond *in-vitro* models, as we perform here, towards true *in-vivo* settings. Integration of these datasets therefore represents a potential path forward for systematic reconstruction of the regulatory factors and networks that govern post-transcriptional regulation and the RNA life cycle.

DATA AND CODE AVAILABILITY

The CPA-Perturb-seq datasets generated for this manuscript are available for download at:
https://zenodo.org/record/7619593#.Y-P7Zi1h2X0

Seurat and PASTA are both available as open-source R packages at:
https://github.com/satijalab/seurat
https://github.com/satijalab/PASTA

Code to train and interpret the APARENT-Perturb model is available at
https://github.com/johli/aparent-perturb

ACKNOWLEDGEMENTS

The authors would like to acknowledge Torben Heick Jensen, Robert Bradley, Christina Leslie, and Christine Mayr for thoughtful discussions related to this work. We thank the Neville Sanjana lab for providing access to the KRAB-dCas9-MeCP2 plasmid. The work was supported by the Chan Zuckerberg Initiative (EOSS5-0000000381, HCA-A-1704- 01895 to R.S.), and the NIH (RM1HG011014-02, 1OT2OD033760-01 to R.S). A.K. and J.L. were supported by NIH grants 2U24HG007234.

COMPETING INTERESTS

In the past three years, R.S. has worked as a consultant for Bristol-Myers Squibb, Regeneron, and Kallyope and served as an SAB member for ImmunAI, Resolve Biosciences, Nanostring, and the NYC Pandemic Response Lab. A.K. is on the SAB of PatchBio Inc., SerImmune Inc., AINovo Inc., TensorBio Inc., and OpenTargets; was a consultant with Illumina Inc. until Jan 2023; and owns shares in DeepGenomics Inc., Immunai Inc. and Freenome Inc. J.L. is an employee of Calico Life Sciences LLC as of 11/21/2022.

REFERENCES

1. Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol. Cell *43*, 853–866.

2. Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30.

3. Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev. *25*, 1770–1782.

4. Gruber, A.J., and Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. Nat. Rev. Genet. *20*, 599–614.

5. Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. *33*, 201–212.

6. Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., and Milos, P.M. (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell *143*, 1018–1029.

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

8. Berkovits, B.D., and Mayr, C. (2015). Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. Nature *522*, 363–367.

9. Arora, A., Goering, R., Lo, H.Y.G., Lo, J., Moffatt, C., and Taliaferro, J.M. (2021). The Role of Alternative Polyadenylation in the Regulation of Subcellular RNA Localization. Front. Genet. *12*, 818668.

10. Lee, S.-H., Singh, I., Tisdale, S., Abdel-Wahab, O., Leslie, C.S., and Mayr, C. (2018). Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. Nature *561*, 127–131.

11. Singh, I., Lee, S.-H., Sperling, A.S., Samur, M.K., Tai, Y.-T., Fulciniti, M., Munshi, N.C., Mayr, C., and Leslie, C.S. (2018). Widespread intronic polyadenylation diversifies immune cell transcriptomes. Nat. Commun. *9*, 1716.

12. Tian, B., Pan, Z., and Lee, J.Y. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res. *17*, 156–165.

13. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science *320*, 1643–1647.

14. Yuan, F., Hankey, W., Wagner, E.J., Li, W., and Wang, Q. (2021). Alternative polyadenylation of mRNA and its role in cancer. Genes Dis *8*, 61–72.

15. Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673–684.

16. Agarwal, V., Lopez-Darwin, S., Kelley, D.R., and Shendure, J. (2021). The landscape of alternative polyadenylation in single cells of the developing mouse embryo. Nat. Commun. *12*, 5101.

17. Cheng, L.C., Zheng, D., Baljinnyam, E., Sun, F., Ogami, K., Yeung, P.L., Hoque, M., Lu, C.-W.,

Manley, J.L., and Tian, B. (2020). Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation. Nat. Commun. *11*, 3182.

18. Leung, M.K.K., Delong, A., and Frey, B.J. (2018). Inference of the human polyadenylation code. Bioinformatics *34*, 2889–2898.

19. Arefeen, A., Xiao, X., and Jiang, T. (2019). DeepPASTA: deep neural network based polyadenylation site analysis. Bioinformatics *35*, 4577–4585.

20. Bogard, N., Linder, J., Rosenberg, A.B., and Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell *178*, 91–106.e23.

21. Li, G.-W., Nan, F., Yuan, G.-H., Liu, C.-X., Liu, X., Chen, L.-L., Tian, B., and Yang, L. (2021). SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells. Genome Biol. *22*, 221.

22. Lianoglou, S., Garg, V., Yang, J.L., Leslie, C.S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. *27*, 2380–2396.

23. Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat. Methods *10*, 133–139.

24. Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W., and Zavolan, M. (2016). A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. Genome Res. *26*, 1145–1159.

25. Brumbaugh, J., Di Stefano, B., Wang, X., Borkent, M., Forouzmand, E., Clowers, K.J., Ji, F., Schwarz, B.A., Kalocsay, M., Elledge, S.J., et al. (2018). Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. Cell *172*, 106–120.e21.

26. Gregersen, L.H., Mitter, R., Ugalde, A.P., Nojima, T., Proudfoot, N.J., Agami, R., Stewart, A., and Svejstrup, J.Q. (2019). SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. Cell *177*, 1797–1813.e18.

27. Schwich, O.D., Blümel, N., Keller, M., Wegener, M., Setty, S.T., Brunstein, M.E., Poser, I., Mozos, I.R.D.L., Suess, B., Münch, C., et al. (2021). SRSF3 and SRSF7 modulate 3'UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels. Genome Biol. *22*, 82.

28. Giammartino, D.C.D., Di Giammartino, D.C., Li, W., Ogami, K., Yashinskie, J.J., Hoque, M., Tian, B., and Manley, J.L. (2014). RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3′ UTRs. Genes & Development *28*, 2248–2260. 10.1101/gad.245787.114.

29. Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L., et al. (2015). Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. PLoS Genet. *11*, e1005166.

30. Ogorodnikov, A., Levin, M., Tattikota, S., Tokalov, S., Hoque, M., Scherzinger, D., Marini, F., Poetsch, A., Binder, H., Macher-Göppinger, S., et al. (2018). Transcriptome 3'end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma.

Nat. Commun. *9*, 5331.

31. Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell *167*, 1883–1896.e15.

32. Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867–1882.e21.

33. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853–1866.e17.

34. Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301.

35. Patrick, R., Humphreys, D.T., Janbandhu, V., Oshlack, A., Ho, J.W.K., Harvey, R.P., and Lo, K.K. (2020). Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. Genome Biol. *21*, 167.

36. Gao, Y., Li, L., Amos, C.I., and Li, W. (2021). Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. Genome Res. *31*, 1856–1866.

37. Fansler, M.M., Zhen, G., and Mayr, C. (2021). Quantification of alternative 3′UTR isoforms from single cell RNA-seq data with scUTRquant. bioRxiv, 2021.11.22.469635. 10.1101/2021.11.22.469635.

38. Harrison, P., Williams, S., Powell, D., Albrecht, D., and Beilharz, T. (2019). Tools for identifying and characterizing alternative polyadenylation in scRNA-Seq. 10.7490/f1000research.1117076.1.

39. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *185*, 2559–2575.e28.

40. Yang, Y., Li, W., Hoque, M., Hou, L., Shen, S., Tian, B., and Dynlacht, B.D. (2016). PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. PLoS Genet. *12*, e1005794.

41. Katahira, J., Okuzaki, D., Inoue, H., Yoneda, Y., Maehara, K., and Ohkawa, Y. (2013). Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. Nucleic Acids Res. *41*, 7060–7072.

42. Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck, W.M., 3rd, Wessels, H.-H., Hao, Y., Yeung, B.Z., Smibert, P., et al. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nat. Genet. *53*, 322–331.

43. Townes, F.W., Hicks, S.C., Aryee, M.J., and Irizarry, R.A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. *20*, 295.

44. Lause, J., Berens, P., and Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. *22*, 258.

45. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell

RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

46. Choudhary, S., and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. *23*, 27.

47. Mosimann, J.E. (1962). On the Compound Multinomial Distribution, the Multivariate β- Distribution, and Correlations Among Proportions. Biometrika *49*, 65–82.

48. McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297.

49. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

50. Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. U. S. A. *99*, 12795–12800.

51. Fatscher, T., Boehm, V., Weiche, B., and Gehring, N.H. (2014). The interaction of cytoplasmic poly(A)-binding protein with eukaryotic initiation factor 4G suppresses nonsense-mediated mRNA decay. RNA *20*, 1579–1592.

52. Hou, L., Wang, Y., Liu, Y., Zhang, N., Shamovsky, I., Nudler, E., Tian, B., and Dynlacht, B.D. (2019). Paf1C regulates RNA polymerase II progression by modulating elongation rate. Proc. Natl. Acad. Sci. U. S. A. *116*, 14583–14592.

53. Pinto, P.A.B., Henriques, T., Freitas, M.O., Martins, T., Domingues, R.G., Wyrzykowska, P.S., Coelho, P.A., Carmo, A.M., Sunkel, C.E., Proudfoot, N.J., et al. (2011). RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J. *30*, 2431–2444.

54. O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front. Endocrinol. *9*, 402.

55. Chen, C.Y., and Shyu, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem. Sci. *20*, 465–470.

56. Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.-B., Li, W., and Wagner, E.J. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature *510*, 412–416.

57. Linder, J., Koplik, S.E., Kundaje, A., and Seelig, G. (2022). Deciphering the impact of genetic variation on human polyadenylation using APARENT2. Genome Biol. *23*, 232.

58. Li, Z., Li, Y., Zhang, B., Li, Y., Long, Y., Zhou, J., Zou, X., Zhang, M., Hu, Y., Chen, W., et al. (2022). DeeReCT-APA: Prediction of Alternative Polyadenylation Site Usage Through Deep Learning. Genomics Proteomics Bioinformatics *20*, 483–495.

59. Vainberg Slutskin, I., Weinberger, A., and Segal, E. (2019). Sequence determinants of polyadenylation-mediated regulation. Genome Res. *29*, 1635–1647.

60. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366.

61. Cheng, J., Çelik, M.H., Kundaje, A., and Gagneur, J. (2021). MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. *22*, 94.

62. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods *12*, 931–934.

63. Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. *28*, 739–750.

64. Shrikumar, A., Tian, K., Shcherbina, A., and Avsec, Ž. Tf-Modisco v0. 4.4. 2-Alpha. arXiv preprint arXiv.

65. Yang, Q., Gilmartin, G.M., and Doublié, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. Proc. Natl. Acad. Sci. U. S. A. *107*, 10062–10067.

66. Yang, Q., Gilmartin, G.M., and Doublié, S. (2011). The structure of human cleavage factor I(m) hints at functions beyond UGUA-specific RNA binding: a role in alternative polyadenylation and a potential link to 5' capping and splicing. RNA Biol. *8*, 748–753.

67. Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

68. Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J.S., Sandelin, A., et al. (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol. Cell *64*, 520–533.

69. Ogami, K., Richard, P., Chen, Y., Hoque, M., Li, W., Moresco, J.J., Yates, J.R., 3rd, Tian, B., and Manley, J.L. (2017). An Mtr4/ZFC3H1 complex facilitates turnover of unstable nuclear RNAs to prevent their cytoplasmic transport and global translational repression. Genes Dev. *31*, 1257–1271.

70. Wu, G., Schmid, M., Rib, L., Polak, P., Meola, N., Sandelin, A., and Jensen, T.H. (2020). A Two-Layered Targeting Mechanism Underlies Nuclear RNA Sorting by the Human Exosome. Cell Rep. *30*, 2387–2401.e5.

71. Gebhardt, A., Habjan, M., Benda, C., Meiler, A., Haas, D.A., Hein, M.Y., Mann, A., Mann, M., Habermann, B., and Pichlmair, A. (2015). mRNA export through an additional cap-binding complex consisting of NCBP1 and NCBP3. Nat. Commun. *6*, 8192.

72. Batra, R., Charizanis, K., Manchanda, M., Mohan, A., Li, M., Finn, D.J., Goodwin, M., Zhang, C., Sobczak, K., Thornton, C.A., et al. (2014). Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. Mol. Cell *56*, 311–322.

73. Itskovich, S.S., Gurunathan, A., Clark, J., Burwinkel, M., Wunderlich, M., Berger, M.R., Kulkarni, A., Chetal, K., Venkatasubramanian, M., Salomonis, N., et al. (2020). MBNL1 regulates essential alternative RNA splicing patterns in MLL-rearranged leukemia. Nat. Commun. *11*, 2369.

74. Arunachalam, P.S., Wimmers, F., Mok, C.K.P., Perera, R.A.P.M., Scott, M., Hagan, T., Sigal, N., Feng, Y., Bristow, L., Tak-Yin Tsang, O., et al. (2020). Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science *369*, 1210–1220.

75. Alt, F.W., Bothwell, A.L., Knapp, M., Siden, E., Mather, E., Koshland, M., and Baltimore, D. (1980). Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. Cell *20*, 293–301.

76. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*,

3573–3587.e29.

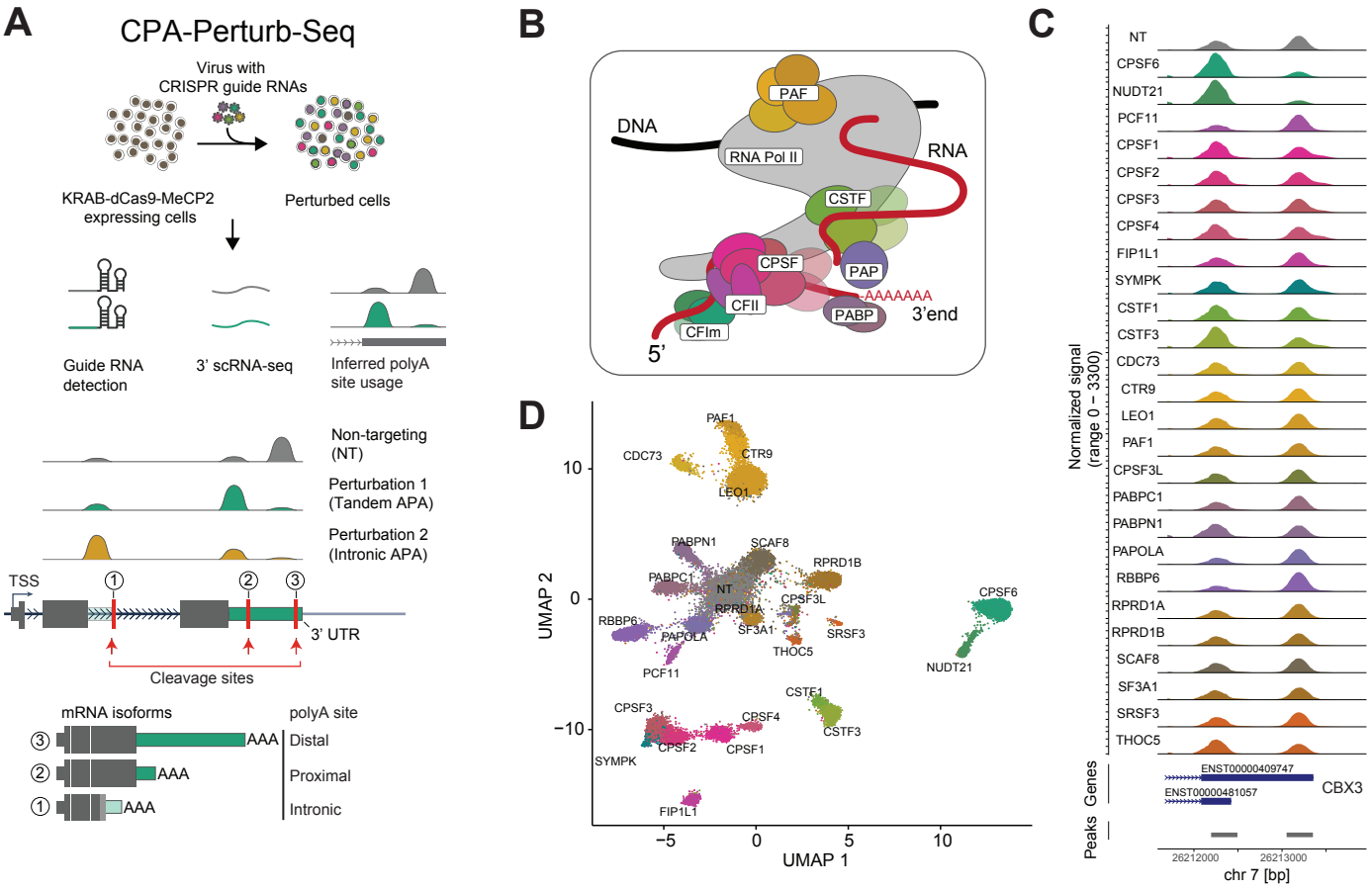
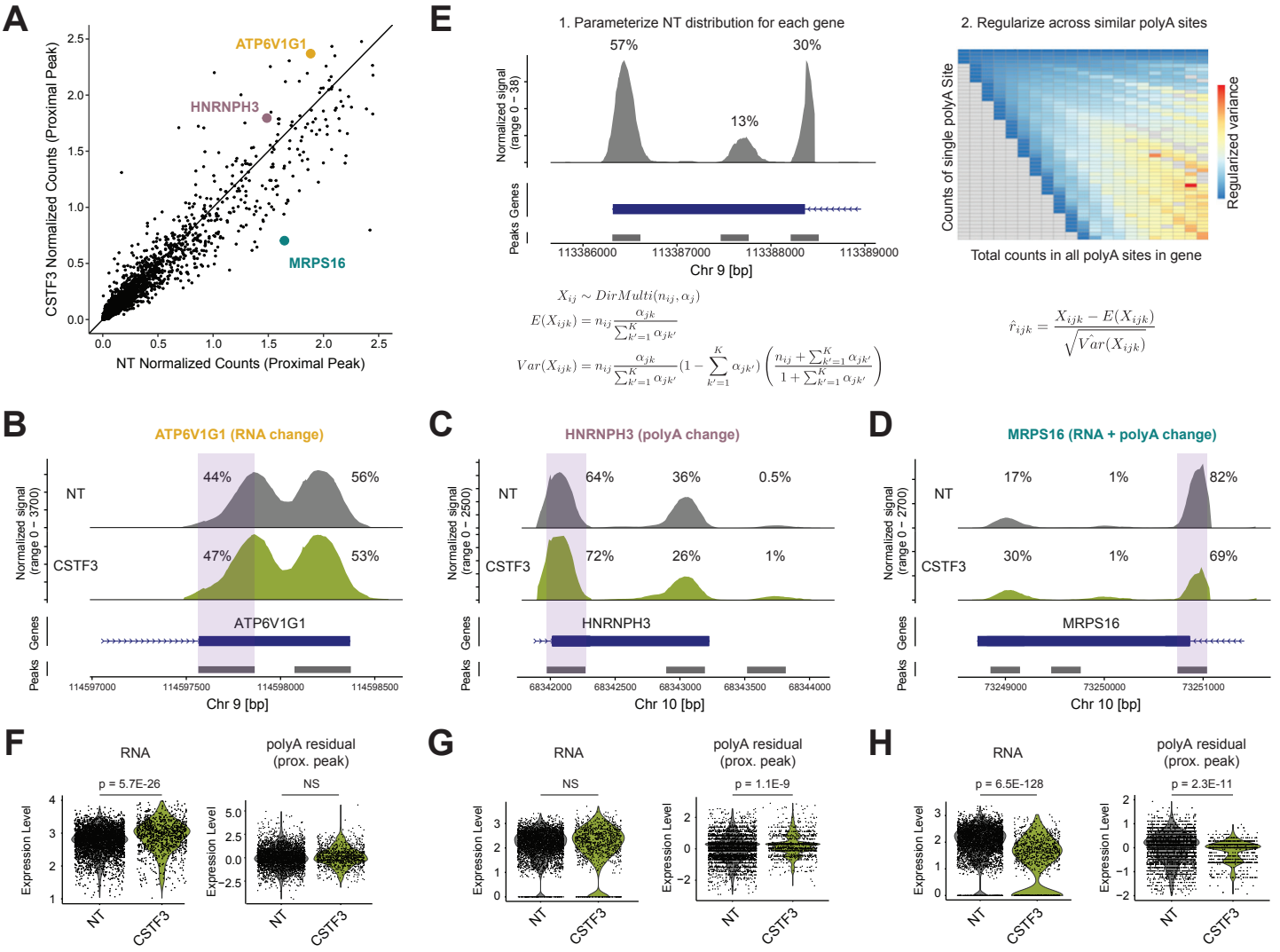


Figure 1: Overview of CPA-Perturb-Seq.
(A) (Top) Schematic of the experimental workflow used to generate the CPA-Perturb-seq dataset. (Bottom) Schematic of perturbation-dependent changes in either tandem or intronic polyadenylation. **(B)** Diagram depicting core regulatory complexes that make up and interact with the cleavage and polyadenylation machinery. **(C)** Read coverage plots depicting the differential use of alternative polyA sites at the CBX3 locus. Each track represents a pseudobulk average of cells, grouped by their perturbation. ENSEMBL gene models and peaks (quantification region) that precede detected polyA sites are shown below. **(D)** UMAP visualization of HEK293FT cells profiled via CPA-Perturb-Seq. Cells are colored based on the target gene identity. Visualization was computed based on a linear discriminant analysis (LDA) of transcriptome-wide polyA site counts.



$$X_{ij} \sim DirMulti(n_{ij}, \alpha_j)$$

$$E(X_{ijk}) = n_{ij}\frac{\alpha_{jk}}{\sum_{k'=1}^{K}\alpha_{jk'}}$$

$$Var(X_{ijk}) = n_{ij}\frac{\alpha_{jk}}{\sum_{k'=1}^{K}\alpha_{jk'}}(1-\sum_{k'=1}^{K}\alpha_{jk'})\left(\frac{n_{ij}+\sum_{k'=1}^{K}\alpha_{jk'}}{1+\sum_{k'=1}^{K}\alpha_{jk'}}\right)$$

$$\hat{r}_{ijk} = \frac{X_{ijk} - E(X_{ijk})}{\sqrt{\hat{Var}(X_{ijk})}}$$

Figure 2: PolyA-residuals quantify alternative polyadenylation at single-cell resolution.
(A) Average usage of 5,335 proximal polyA sites in NT cells (x-axis) and CSTF3-perturbed cells (y-axis). Only genes with at least two tandem polyA sites are considered. Changes across conditions can breflect eithevr changes in relative polyA site usage, total gene expression, or both**. (B-D):** Read coverage plots at three loci highlighted in (A). Blue box denotes proximal polyA site. **(E)** Schematic depicting the procedure to calculate polyA-residuals (full description in Supplementary Methods). **(F-H)** Violin plots depicting single-cell gene expression levels (left) or single-cell polyA-residuals for the proximal polyA site for NT and CSTF3-perturbed cells. NS (not significant) for RNA comparisons indicate absolute log2FC <0.25 or Bonferonni adjusted p-value >0.05 using Wilcoxon rank-sum test. NS for polyA residual comparisons indicates percent change <0.05 or adjusted p-value >0.05 in differential polyadenylation analysis described in Supplementary Methods.

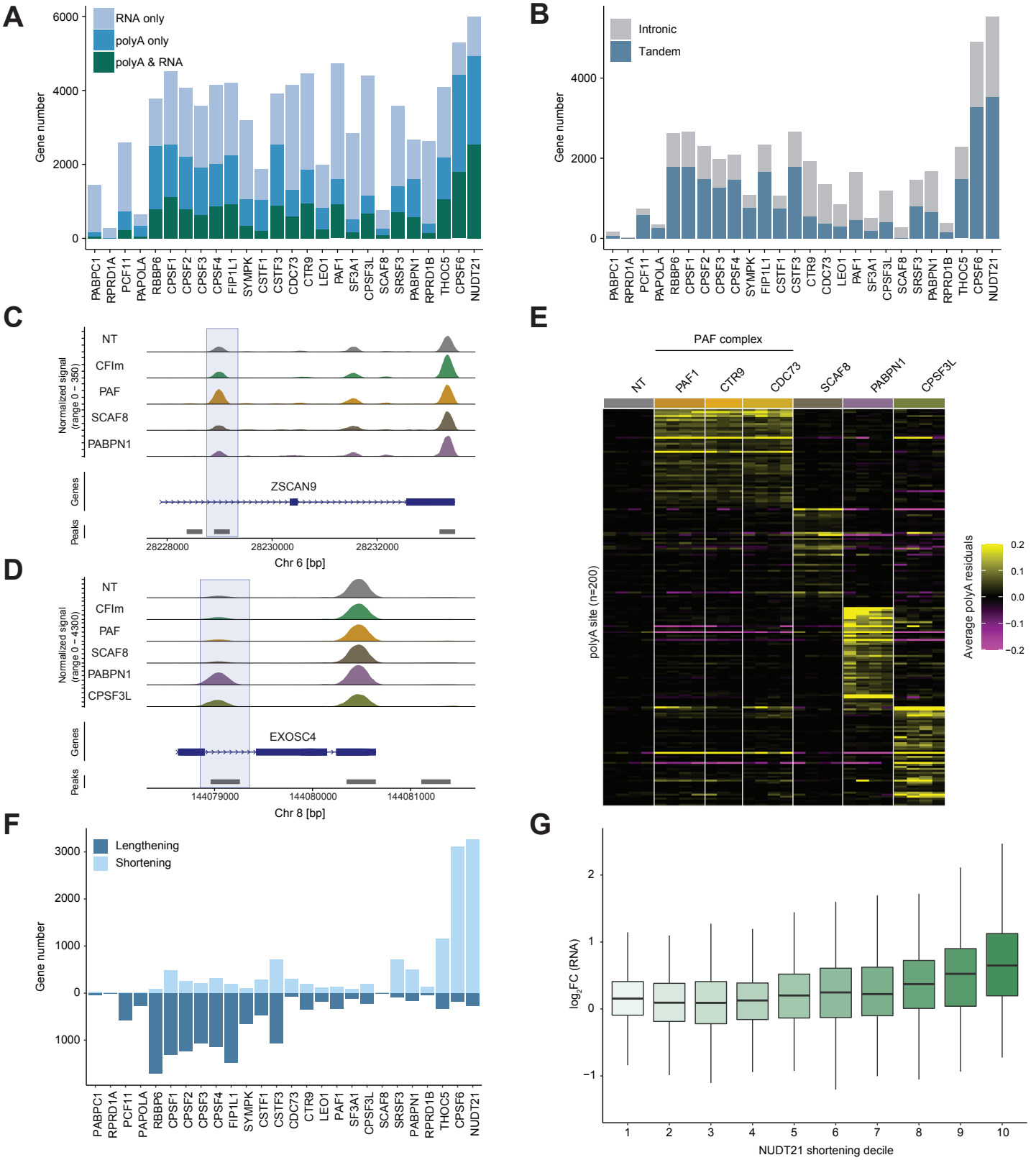


Figure 3: Characterizing tandem and intronic alternative polyadenylation in CPA-Perturb-Seq
(A) Number of genes with significant changes in RNA abundance, relative usage of at least one polyA site, after perturbation of each regulator. Barplots show results in HEK293FT cells. **(B)** Total number of genes with relative changes in intronic or tandem polyA site usage in HEK293FT cell dataset. **(C-D)** Read coverage plots showing differential usage of intronic sites (boxed) at the ZSCA9 (C) and EXOSC4 locus (D). **(E)** Heatmap showing polyA residuals for intronic sites that are uniquely differentially utilized after perturbation of PAF, SCAF8, PABPN1, and CPSF3L. Each heatmap cell shows the pseudobulk average of cells after grouping by sgRNA identity. **(F)** Number of genes with significant changes in tandem polyA site usage in HEK293FT cell dataset, classified by 3' UTR shortening or 3' UTR lengthening. **(G)** Boxplot indicating the observed log2 fold-change in gene expression after NUDT21 perturbation. Genes are partitioned into deciles based on the degree of 3' UTR changes observed after NUDT21 perturbation.

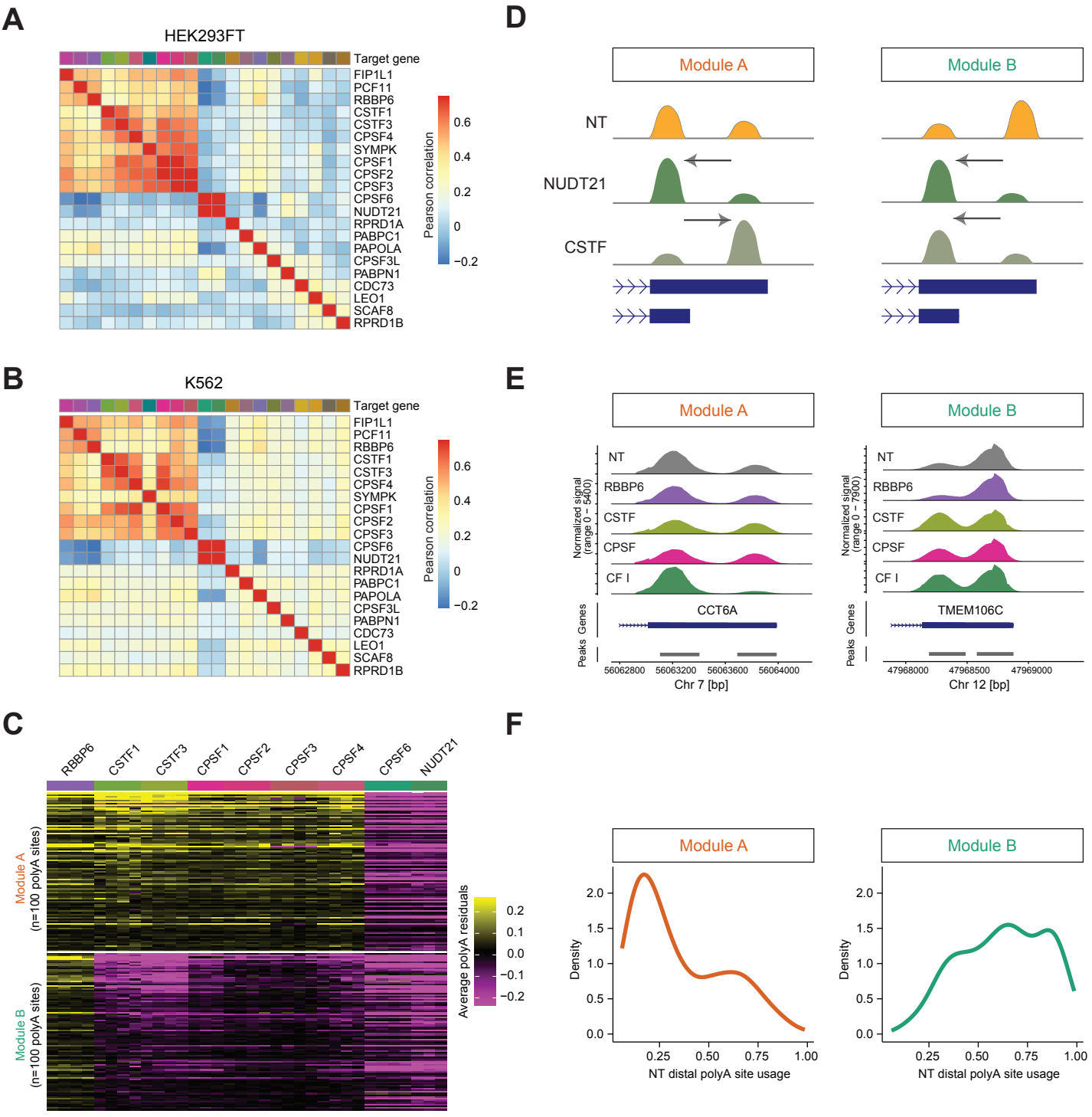


Figure 4: Modules of co-regulated polyA sites exhibit functional differences
(A) Pearson correlation matrix depicting the relationships between perturbations in HEK293FT cells. Correlations are calculated using the linear model coefficients learned during differential polyadenylation analysis (Supplementary Methods). Matrices include all perturbations where we obtained at least 50 cells in both HEK and K562 cells, and are ordered via hierarchical clustering. **(B)** Same as (A), but the correlation matrix is generated from an independent analysis on K562 polyA residuals. **(C)** Heatmap showing polyA-residuals for distal peak sites in Module A genes (CSTF and CPSF act in the opposite direction as CPSF6/NUDT21), and Module B genes (CSTF and CPSF act in the same direction as CPSF6/NUDT21). For visualization, the top 100 polyA sites, ranked by the magnitude of CSTF perturbation, are shown for each module. **(D)** Schematic diagram of genes belonging to module A and Module B. **(E)** Read coverage plots showing polyA site usage of representative genes belonging to module A (left, CCT6A) and module B (right, TMEM106C). **(F)** Density plot showing distal site usage in NT control cells for genes belonging to Module A (left) versus Module B (right). Genes in Module A tend to use the proximal site, while genes in Module B tend to use the distal site.

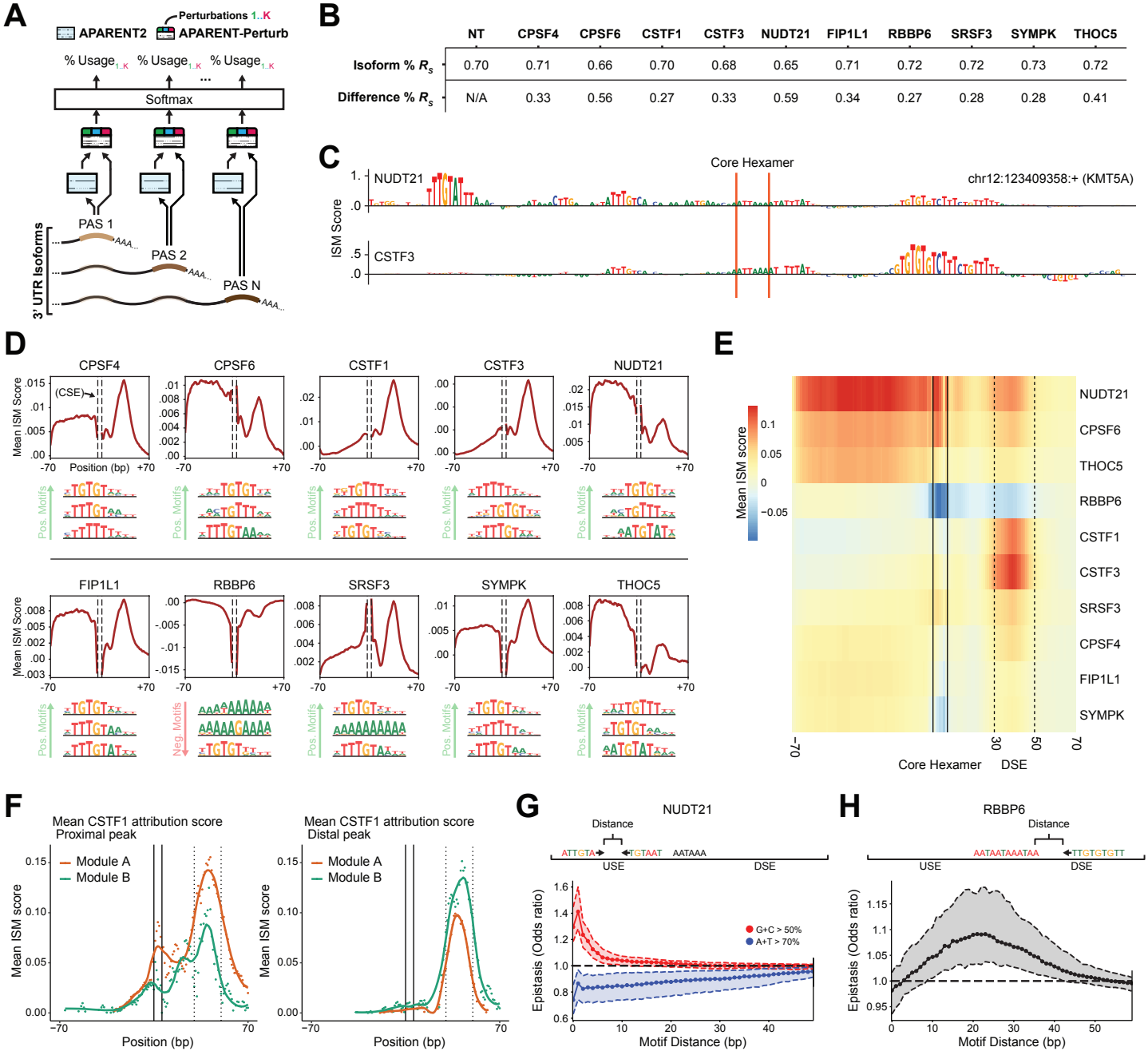


Figure 5. A multi-task neural network predicts perturbation responses from RNA sequence.
(A) Schematic of (APARENT-Perturb), an ensemble-based neural network architecture for predicting perturbation responses. Green/blue/red output heads correspond to model predictions for the K perturbation conditions. **(B)** 10-fold cross-validation performance when predicting distal isoform proportions (top row) or differences in distal isoform proportion with respect to the NT condition (bottom row). **(C)** Sequence-specific attribution scores for two example perturbations in the KMT5A gene. Attribution scores are displayed after calculating residuals with respect to NT cells. **(D)** Averaged attribution scores as a function of position, for 10 perturbations. The three top MoDISco motifs are shown for each perturbation (Supplementary Methods). **(E)** Heatmap showing averaged attribution scores for each perturbation, as a function of position, for the distal-most site in each gene. **(F)** Mean attribution scores for CSTF1 perturbation in Module A vs Module B for both proximal sites (left) and distal sites (right). Location of the core hexamer and downstream sequence elements (DSE) are marked with solid and dashed vertical lines, respectively. Plots show the mean attribution score at single base-pair resolution (points), as well as the loess-smoothed trend (lines). **(G)** Epistasis analysis for dual UGUA motifs, in either G/C-rich contexts (red) or A/U-rich contexts (blue). The y-axis reflects the effect on predicted NUDT21 perturbation after dual insertion of both motifs, compared to the effect of inserting one motif at a time (Supplementary Methods). **(H)** Epistasis analysis of canonical hexamers and U/G-rich motifs, based on the RBBP6 perturbation.

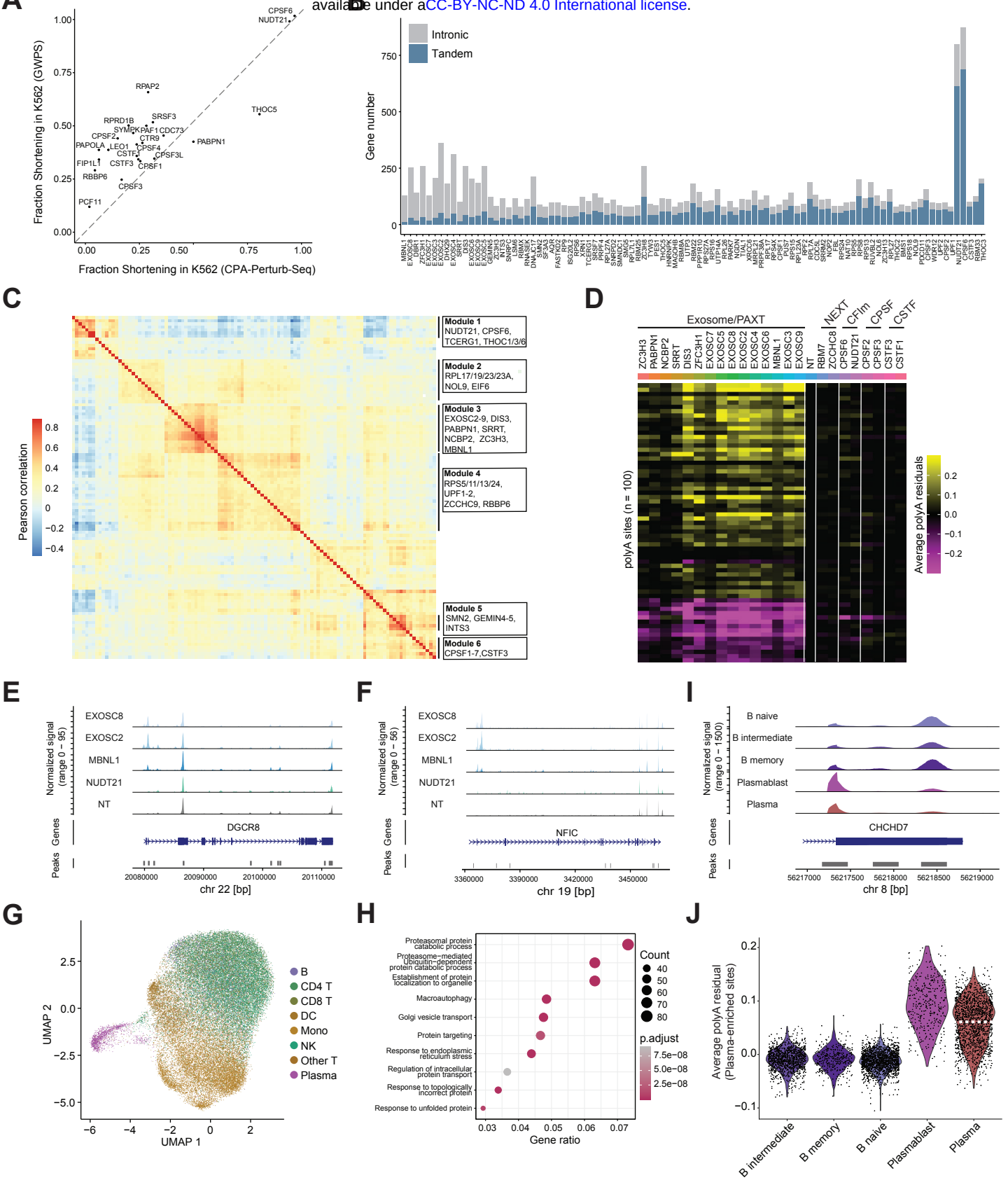


Figure 6: Characterizing heterogeneity in relative polyA site usage in additional 3' scRNA-seq datasets

(A) 3' UTR shortening preference observed after perturbing regulators in the CPA-Perturb-seq dataset (x-axis), and the GWPS dataset (y-axis). We observe concordant results for this global metric across datasets. **(B)** Same as Figure 3B, but for the GWPS dataset. **(C)** Correlation matrix depicting the relationship between perturbations in the GWPS dataset, as in Figure 4A. Representative genes for each of the six correlated modules are shown on the left. All genes are listed in Supplementary Figure 6A. Shown are all perturbations where we detected changes in relative polyA site usage in at least 50 genes. **(D-F)** MBNL1 perturbation phenocopies perturbation of PAXT complex members. **(D)** Heatmap shows polyA-residuals for polyA sites that are differentially utilized after both MBNL1 and PAXT perturbation. **(E-F)** Representative read coverage plot depicting changes in polyA site usage after perturbation of PAXT complex members and MBNL1. **(G)** UMAP visualization generated from polyA residuals of PBMC dataset. Cells are colored based on their gene expression-based cell annotation. **(H)** Gene ontology enrichment analysis on genes exhibiting 3'UTR shortening in plasma cells compared to B cells. **(I)** Read coverage plot depicting 3'UTR shortening in CHCHD7 gene in distinct B and plasma cell subpopulations. **(J)** Average polyA-residual (reflects degree of 3' UTR shortening) of proximal sites with increased usage in plasma cells. We observe extensive heterogeneity within the plasma cell lineage, including increased shortening in cycling plasmablasts, and two subpopulations of non-cycling plasma cells (denoted by horizontal line, Supplementary Figure 7C).