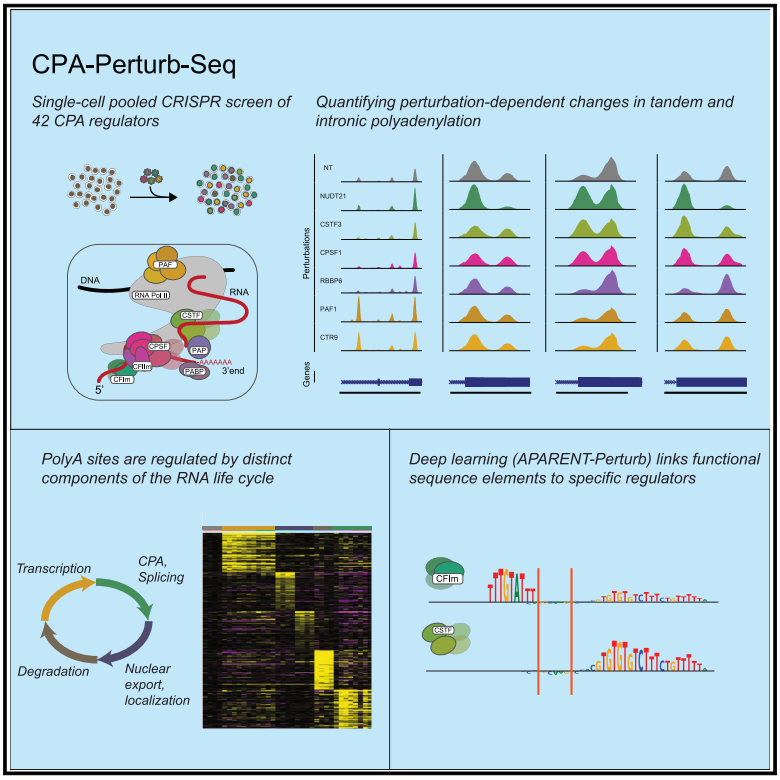
Cell

Resource

Multiplexed single-cell characterization of alternative polyadenylation regulators

Graphical abstract



Highlights

- Genetic perturbations of CPA regulators reveal coordinated changes in polyA site usage
- PASTA quantifies heterogeneity in polyA site usage in scRNA-seq data
- Distinct components of RNA life cycle affect intronic polyA site usage
- Deep learning identifies sequence features that dictate perturbation response

Authors

Madeline H. Kowalski, Hans-Hermann Wessels, Johannes Linder, ..., Yuhan Hao, Anshul Kundaje, Rahul Satija

Correspondence

wesselshanshermann@gmail.com (H.-H.W.), rsatija@nygenome.org (R.S.)

In brief

CRISPR perturbation screens of 42 cleavage and polyadenylation regulators followed by 3′ scRNA-seq (CPA-Perturb-seq) identifies modules of co-regulated polyA sites, their connection with distinct components of the RNA life cycle, and the sequence determinants that drive perturbation responses.

Kowalski et al., 2024, Cell *187*, 4408–4425
August 8, 2024
https://doi.org/10.1016/j.cell.2024.06.005

CellPress

CellPress

Cell

Resource

Multiplexed single-cell characterization of alternative polyadenylation regulators

Madeline H. Kowalski,[1,2,3,6] Hans-Hermann Wessels,[1,2,6,*] Johannes Linder,[4,5,6] Carol Dalgarno,[1] Isabella Mascio,[1,2] Saket Choudhary,[1,2] Austin Hartman,[1] Yuhan Hao,[1,2] Anshul Kundaje,[4,5] and Rahul Satija[1,2,3,7,*]

[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, CA, USA
[5]Department of Computer Science, Stanford University, Stanford, CA, USA
[6]These authors contributed equally
[7]Lead contact
*Correspondence: wesselshanshermann@gmail.com (H.-H.W.), rsatija@nygenome.org (R.S.)
https://doi.org/10.1016/j.cell.2024.06.005

SUMMARY

Most mammalian genes have multiple polyA sites, representing a substantial source of transcript diversity regulated by the cleavage and polyadenylation (CPA) machinery. To better understand how these proteins govern polyA site choice, we introduce CPA-Perturb-seq, a multiplexed perturbation screen dataset of 42 CPA regulators with a 3′ scRNA-seq readout that enables transcriptome-wide inference of polyA site usage. We develop a framework to detect perturbation-dependent changes in polyadenylation and characterize modules of co-regulated polyA sites. We find groups of intronic polyA sites regulated by distinct components of the nuclear RNA life cycle, including elongation, splicing, termination, and surveillance. We train and validate a deep neural network (APARENT-Perturb) for tandem polyA site usage, delineating a *cis*-regulatory code that predicts perturbation response and reveals interactions between regulatory complexes. Our work highlights the potential for multiplexed single-cell perturbation screens to further our understanding of post-transcriptional regulation.

INTRODUCTION

RNA cleavage and polyadenylation (CPA) 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 can influence multiple distinct stages of the RNA life cycle. For example, shortening of the 3′ untranslated region (UTR) can affect transcript stability and localization,[8,9] while alternative polyadenylation at intronic sites can lead to the generation of truncated coding or noncoding 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 CPA specificity factor (CPSF) complex catalyzes cleavage, the cleavage factor Im (CFIm) and cleavage factor IIm (CFIIm) complexes bind auxiliary recognition sequences, and polyA polymerase is responsible for adding the polyA tail.[4]

Genome-wide 3′ transcriptome technologies can be used to profile changes in polyA site usage,[16,18–20] including genetic perturbations of CPA regulators. While some studies perform individual or small sets of perturbations,[21–24] others have used small-interfering-RNA-based screening approaches to generate larger resources.[25,26] These studies have characterized the global tendencies of individual regulators to promote either proximal or distal polyA site usage within the 3′ UTR, regulating a process called tandem alternative polyadenylation (Figure 1A). While different perturbations affect different numbers of sites, it is unclear whether this variation reflects functional co-regulation: i.e., whether groups of polyA sites are uniquely sensitive to perturbation of different regulators or subcomplexes. If so, identifying these modules of co-regulated sites and the molecular features driving their usage represents a key goal for enhancing our understanding of CPA regulation.

Alternative polyadenylation can also occur at intronic polyA sites (intronic alternative polyadenylation; Figure 1A), which, in contrast to 3′ UTR changes, results in alterations to the coding sequence. The usage of intronic polyA sites has been linked to multiple regulatory proteins that govern the synthesis or nuclear processing of RNA transcripts.[12,27–30] While changes in intronic polyadenylation have been identified in disease states,[10,11,31] it

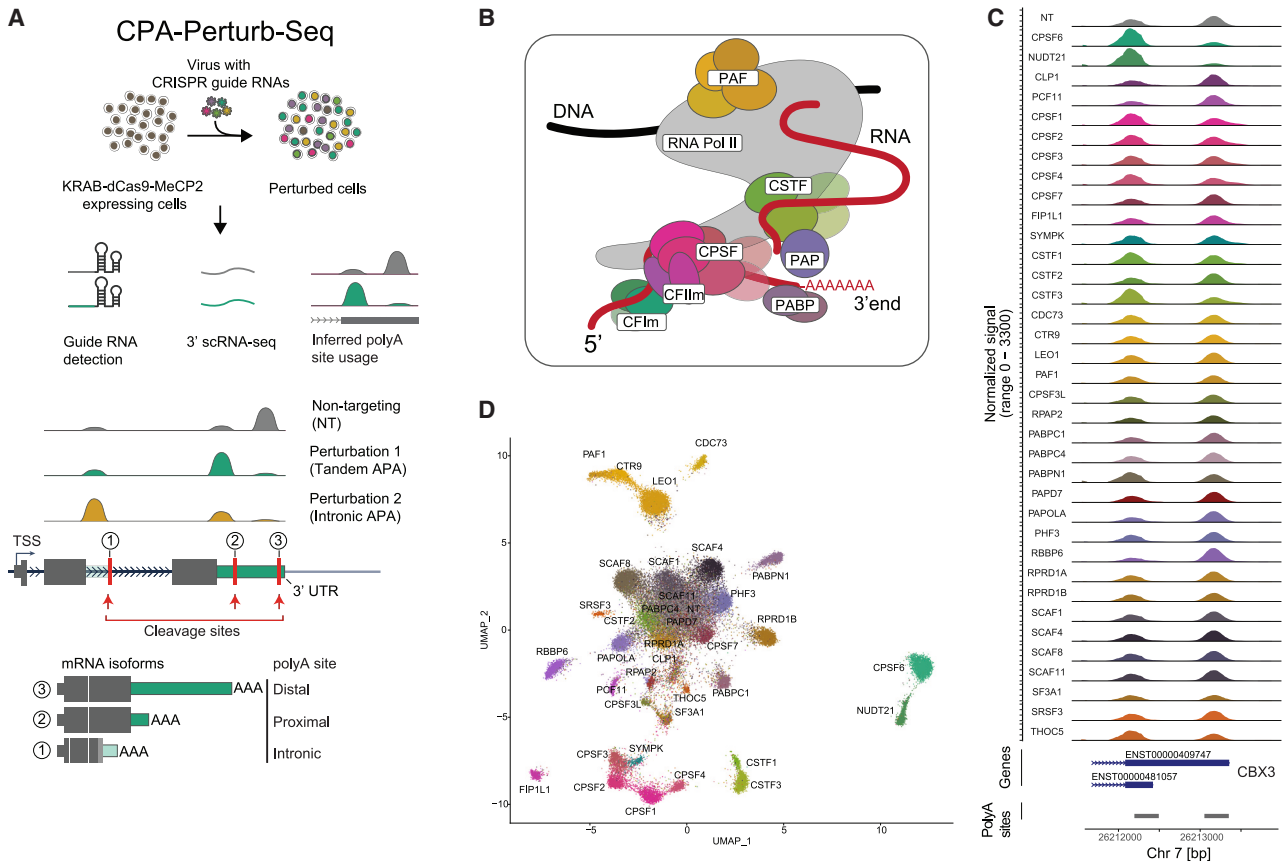


Figure 1. Overview of CPA-Perturb-seq

(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, using the same colors as in (C). Visualization was computed based on a linear discriminant analysis (LDA) of transcriptome-wide polyA site counts.

See also Figures S1–S3.

is unclear if intronic polyA sites are globally sensitive to changes in transcriptional dynamics or if, alternatively, subsets of regulators determine the usage of distinct groups of intronic sites.

Multiplexed single-cell technologies like Perturb-seq, which leverages single-cell RNA sequencing (scRNA-seq) for high-throughput transcriptome-wide characterization of molecular perturbations,[32–36] offer exciting potential to address these questions. 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. The majority of scRNA-seq protocols are 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.[37–40]

Here, we introduce 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 statistical methods to quantify changes in polyA site usage in sparse single-cell datasets, and we identify distinct modules of co-regulated polyA sites. We find that co-regulation of intronic polyA site use is driven by differential sensitivity to perturbation of distinct elements of the nuclear RNA life cycle, while for tandem sites, co-regulation is driven by a *cis*-regulatory code where individual sequence elements modulate responsiveness. We learn this code by extending pioneering deep learning models of alternative polyadenylation[41–46] to multiple genetic contexts, and we validate our findings using a massively parallel reporter assay (MPRA). Finally, we demonstrate how our computational tools can be applied to any 3′ scRNA-seq dataset and characterize the regulatory effects of hundreds of genes involved in RNA processing using a genome-scale Perturb-seq resource.[47] 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 CPA 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 (Table S1). Our target set included 18 genes that are known members of core CPA complexes, including the CFIm, CFIIm, CPSF, and cleavage stimulation factor (CSTF) complexes (Figure 1B). We also included 23 genes that have been previously implicated in affecting relative polyA site usage (Table S1).

We performed a pooled CRISPR interference (CRISPRi) screen in HEK293FT cells and used the Perturb-seq experimental workflow (STAR Methods) to simultaneously capture the identity of the guide each cell received along with a 3′ scRNA-seq readout (Figures 1A and S1). We focused our primary analyses on the deeply profiled HEK293FT dataset (median of 2,168 cells per perturbation) but also repeated the experiment in K562 cells (median of 960 cells per perturbation). Across two biological replicates (independent viral transductions) in each cell line, we obtained a total of 140,415 single cells (Table S2) where we successfully assigned one sgRNA. We applied our previously developed computational pipeline, Mixscape,[48] to address confounding sources of variation that have been described in pooled single-cell CRISPR screens.[33,34,48] For 6 of 42 regulators, Mixscape classified all cells as "non-perturbed," suggesting minimal effects on the transcriptome despite knockdown (KD) of the target gene (Figure S2A). For the remaining 36 genes, Mixscape classified 69% of cells as perturbed.

We utilized the scRNA-seq data to quantify both gene expression and transcriptome-wide polyA site usage profiles for each cell (STAR Methods). We used polyApipe to first identify a set of possible polyA sites and then quantify their usage in single cells, generating a polyA site/cell count matrix for downstream analysis.[40] We restricted our analysis to polyA sites within 50 nucleotides of polyA sites identified in polyA_DB v3.2,[7] a database of polyA sites generated from multiple human cell lines. We also only included sites within an intron or the last exon of a gene (STAR Methods).

We identified a total of 35,882 polyA sites across 12,617 detected genes. We found that 8,558 genes exhibited usage of two or more polyA sites in our dataset (and 5,661 genes exhibited usage of three or more) (Figure S1A). The majority of polyA sites contained the canonical AATAAA/ATTAAA cleavage motif upstream of the cleavage site, as expected (Figure S1B). We validated our predicted polyA sites using 3′ RACE in combination with Illumina sequencing at 7 loci and found that our predicted polyA sites were highly consistent with sensitive mapping of 3′ transcript ends (STAR Methods; Table S3; Figures S1C and S1D).[49,50] We repeated our 3′ RACE experiment in NUDT21-perturbed cells to assess our ability to detect changes in polyA site usage and observed high concordance (R = 0.86) when comparing the effect of NUDT21 perturbation between the 3′ RACE and CPA-Perturb-seq data (Figures S1E–S1G). We also found high correlation between our estimated polyA site usage ratios and transcriptome-wide quantifications derived using bulk gold-standard assays PAPERCLIP and A-seq (STAR Methods; Figures S1H and S1I).[51,52]

We observed diverse effects on polyA site usage across regulators (Figure 1C and S3), and these changes were reproducible across biological replicates and multiple independent sgRNAs (3–4 per gene; Figure S2B). We used the polyA site/cell count matrix for perturbed cells, along with 4,336 NT controls, as input to linear discriminant analysis (LDA), UMAP visualization (Figure 1D), and unsupervised clustering of the polyA site matrix (Figure S2C). These analyses revealed that cells clustered not only by perturbation but also into broader complexes. For example, cellular profiles after NUDT21 and CPSF6 (both members of the CFIm complex) perturbation were highly correlated, as were profiles for members of the CPSF (CPSF1–4 and FIP1L1), CSTF (CSTF1/3), and polymerase-associated factor (PAF) (PAF1, CTR9, LEO1, and 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 functionally related regulators. However, we note that changes in the polyA site/cell count matrix can reflect both changes in polyA site utilization and changes in the overall abundance of the gene. For example, when knocking down CSTF3 (Figures 2A–2D), we identify cases where changes in the utilization of a gene's proximal polyA site correspond exclusively to a change in total RNA abundance (ATP6V1G1), exclusively to a change in transcript length due to 3′ UTR shortening (HNRNPH3), or to changes in both abundance and relative isoform usage (CDK1).

Quantifying relative polyadenylation levels at single-cell resolution

To specifically characterize perturbation-driven effects on alternative polyadenylation, we 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. We clearly observed usage of multiple polyA sites within the same cell, but single-cell polyA site usage ratios were noisy at lower sequencing depths (Figures S2D and S2E). Instead, for each polyA site in each single cell, we aimed to model the degree of over- or underutilization compared with control cells.

We note that this problem is conceptually similar to quantifying changes in gene expression in individual cells, as we and others have addressed using generalized linear models.[53–56] We extended this framework to model alternative polyadenylation (Figure 2E). We utilized the Dirichlet-multinomial distribution to model a background distribution of polyA site usage in NT cells, controlling for gene expression. The Dirichlet multinomial allows for overdispersion compared with the standard multinomial,[57] analogous to use of the negative binomial distribution to model Poisson overdispersion when modeling gene abundances.[58,59] This overdispersion accounts for natural biological heterogeneity

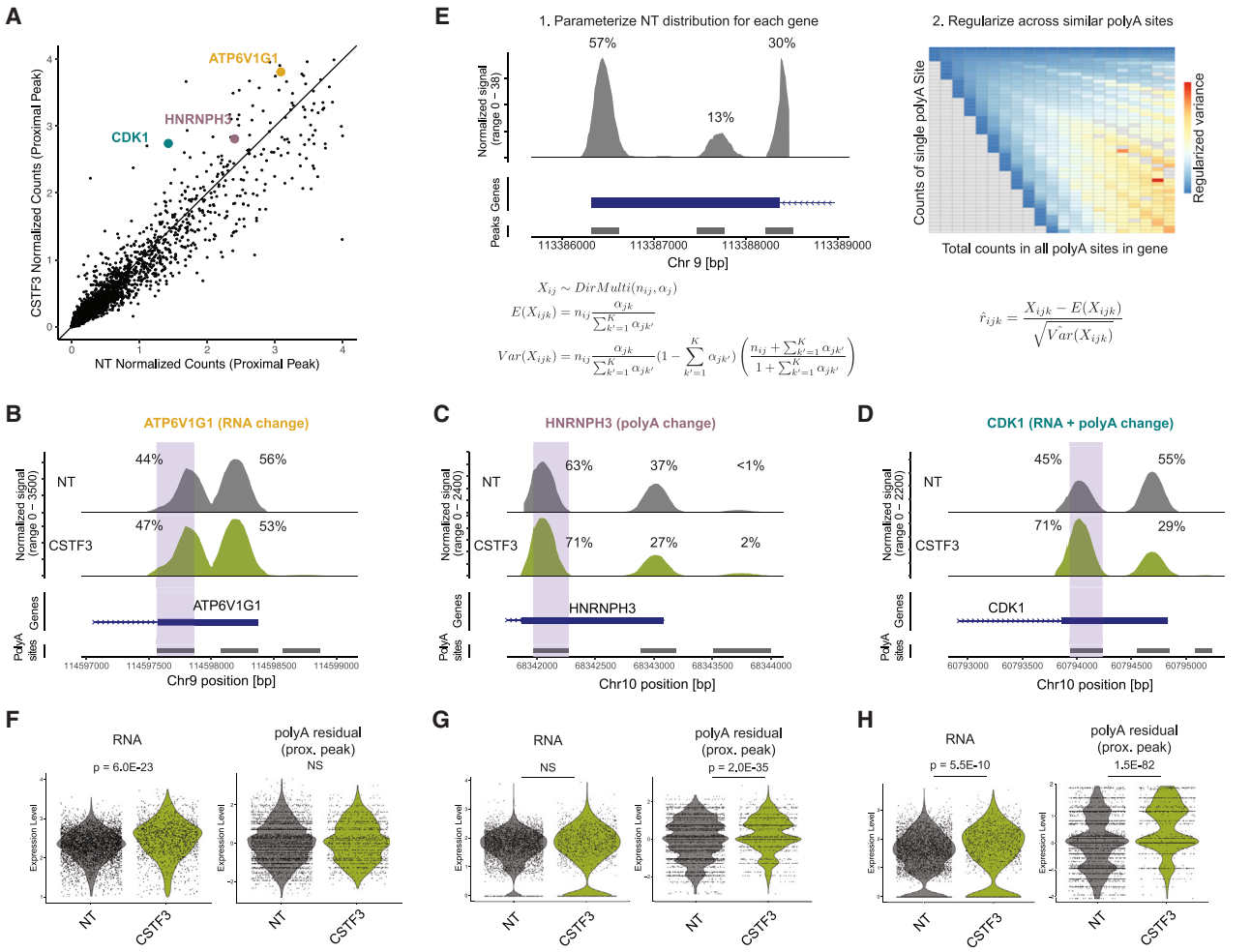


Figure 2. PolyA residuals quantify alternative polyadenylation at single-cell resolution

(A) Average usage of 6,019 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 reflect changes in relative polyA site usage, total gene expression, or both.

(B–D) Read coverage plots at three loci highlighted in (A). Shading marks the proximal polyA site.

(E) Schematic depicting the procedure to calculate polyA residuals (full description in STAR 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. Not significant (NS) for RNA comparisons indicates absolute $\log_2$FC <0.25 or Bonferroni 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 STAR Methods.

and "intrinsic" noise that occurs within the background population.[56,60] As in sctransform,[55] we first parameterize overdispersion estimates individually for each polyA site but then regularize these estimates across similar sites (STAR Methods). The output of our procedure is a statistical model for each polyA site, describing its background usage across 4,336 NT control cells.

By comparing the observed counts at each polyA 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 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 vice versa. We tested for changes in polyA residuals using a linear model, enabling us to identify perturbation-associated changes in polyA site usage without confounding changes in gene abundance (STAR Methods). When applied to our previous examples (Figures 2F–2H), this approach successfully distinguished loci where we observed changes in transcript structure, abundance, or both. Using the polyA-residual matrix for LDA-based visualization (Figure S4A) confirmed that our observed co-regulatory patterns were driven by coordinated changes in polyadenylation.

Characterizing perturbation-dependent changes in polyadenylation

We identified 7,402 genes that exhibited differential alternative polyadenylation (at least one polyA site with differential usage) in at least one of the 36 gene perturbations, but we observed substantial differences across regulators. CFIm complex

members such as NUDT21 exhibited the strongest perturbation responses (Figure 3A), including widespread changes in both polyA site usage and total transcript abundance. By contrast, perturbation of PABPC1 (which binds the polyA tail after nuclear export[61]) primarily affected changes in transcript levels but not structure (Figures 3A and S4B).

We next classified changes in polyA site usage as reflecting either intronic polyadenylation or tandem polyadenylation, and for tandem sites, we determined whether they represented increased usage or the proximal (shortening) or distal (lengthening) site (STAR Methods; Figure 3A). Most regulators affected tandem polyadenylation and exhibited a skew of greater than 70% toward shortening or lengthening, which also replicated in our K562 dataset (Figure S4B; Table S4). 3′ UTR shortening was generally associated with an increase in total gene abundance (Figure S4C), consistent with the association between 3′ UTR length and the presence of regulatory elements that may impact RNA stability.[63,64]

Comparing our results to previous studies that utilize bulk 3′ end sequencing technologies highlighted the advantages of the Perturb-seq technology in defining perturbation signatures. Previous studies[21,25,26,65] have consistently revealed that NUDT21 perturbation affects polyA site usage in a subset of genes (ranging from 375 to 1,600) and leads to 3′ UTR shortening at tandem UTRs. In both our HEK293FT and K562 datasets (Figures 3A and S4B, respectively), we observed high sensitivity (more than 5,400 genes exhibited significant changes in polyA site usage after perturbation) and also high specificity (>91% of tandem UTR changes resulted in shortening). Similarly, RBBP6 perturbation has been associated with 3′ UTR lengthening,[24–26] but our dataset identifies more genes with high specificity (>94% 3′ UTR shortening, Figures 3A and S4B).

Distinct regulatory mechanisms drive intronic polyA site usage

We identified eight regulators where perturbation responses were primarily associated with increased intronic polyadenylation (Figures 3A and S4D). These included anti-terminator proteins SCAF8 and SCAF4,[22] PAF complex members[66] (PAF1, CDC73, and CTR9), PABPN1, RPAP2, and CPSF3L, a member of the integrator complex.[67] Each of these regulators directly interact with RNA polymerase II or newly synthesized transcripts,[68–72] yet perturbation of these regulators affected distinct intronic sites. This suggests multiple modes of regulation of intronic polyadenylation (Figures 3B–3D).

We performed hierarchical clustering to identify intronic regulators with correlated perturbation responses (Figure S4E). We found that PAF complex members[66] (PAF1, CDC73, and CTR9) regulate similar intronic polyA sites (Figure 3D) and that there are similar perturbation responses between integrator complex member CPSF3L and associated protein RPAP2.[73] We performed linear modeling to identify features that predicted changes in intronic polyA site usage (Figure 3E) and found that features' predictive strength varied substantially across regulators. Perturbation of PABPN1 was strongly associated with increased usage of polyA sites within the first intron (Figure 3F). Consistent with this finding, PABPN1 functions as a tandem 3′ UTR regulator[74,75] (Figure 3A), but it also participates in the nuclear surveillance and degradation of short polyadenylated transcripts by associating with factors that recognize the 5′ cap.[68]

We found that intronic GC content and the distance to the next cleavage site in a transcript (a measure of elongation time between cleavage events) were both predictive of the responsiveness to PAF1 perturbation. This relationship was strongest for polyA sites located in the first intron[76] but also held for downstream sites (Figures S4F and S4G). Only 21% of the regulated intronic polyA sites in our dataset were responsive to PAF1 perturbation, highlighting that this specific subset of sites was responsive to broad changes in elongation dynamics.

The integrator subunit CPSF3L and its associated factor RPAP2 regulated intronic polyA sites across the gene body (Figure 3F). These responsive sites were primarily located in short introns (median length: 2,660 bp) and in introns with elevated GC content (Figures 3G, 3H, S4H, and S4I). The integrator complex has been associated with multiple distinct functions, including being involved with small nuclear RNA (snRNA) biogenesis[67] and driving premature termination upon recognizing paused promoter-proximal RNAPII.[72] As we observed no enrichment for introns near transcription start sites (Figure 3F), we instead considered that defects in snRNA processing may drive broader splicing aberrations affecting intronic polyadenylation. We utilized a previously published RNA metabolism dataset that calculated excision speeds for introns from 2,212 genes present in our dataset (Figure 3I).[62] We found that the CPSF3L/RPAP2 signature was enriched for introns with "slow" excision dynamics, which, along with their heightened GC content,[77] suggests that they are inefficiently spliced.

Lastly, we identified a clear bifurcation in the identity of responsive intronic polyA sites from perturbation of two anti-terminator proteins. SCAF4 predominantly regulated polyA site use within short introns (median length: 7,529 bp) with high GC content, while SCAF8 regulated site usage within long introns (median length: 33,798 bp) with low GC content (Figures 3G and 3H). Our findings extend previous pioneering work that suggests that SCAF4 and SCAF8 work redundantly to prevent the use of intronic polyA sites.[22] The sensitivity of our assay allows us to detect phenotypes for both individual perturbations, demonstrates nonredundant roles for these proteins at hundreds of sites, and identifies determinants that guide this selectivity. Our analyses demonstrate that intronic polyadenylation is not a globally regulated phenomenon and that distinct sets of sites are uniquely sensitive to perturbation of factors regulating distinct RNA nuclear life cycle components.

Modules of co-regulated tandem polyA sites exhibit distinct functional properties

We repeated our hierarchical clustering analysis on tandem polyadenylation regulators to identify correlated perturbation responses (STAR Methods; Figure 4A). 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 their perturbation causes 3′ UTR lengthening at overlapping sites. We observed highly concordant correlations in our K562 dataset (Figure 4B).

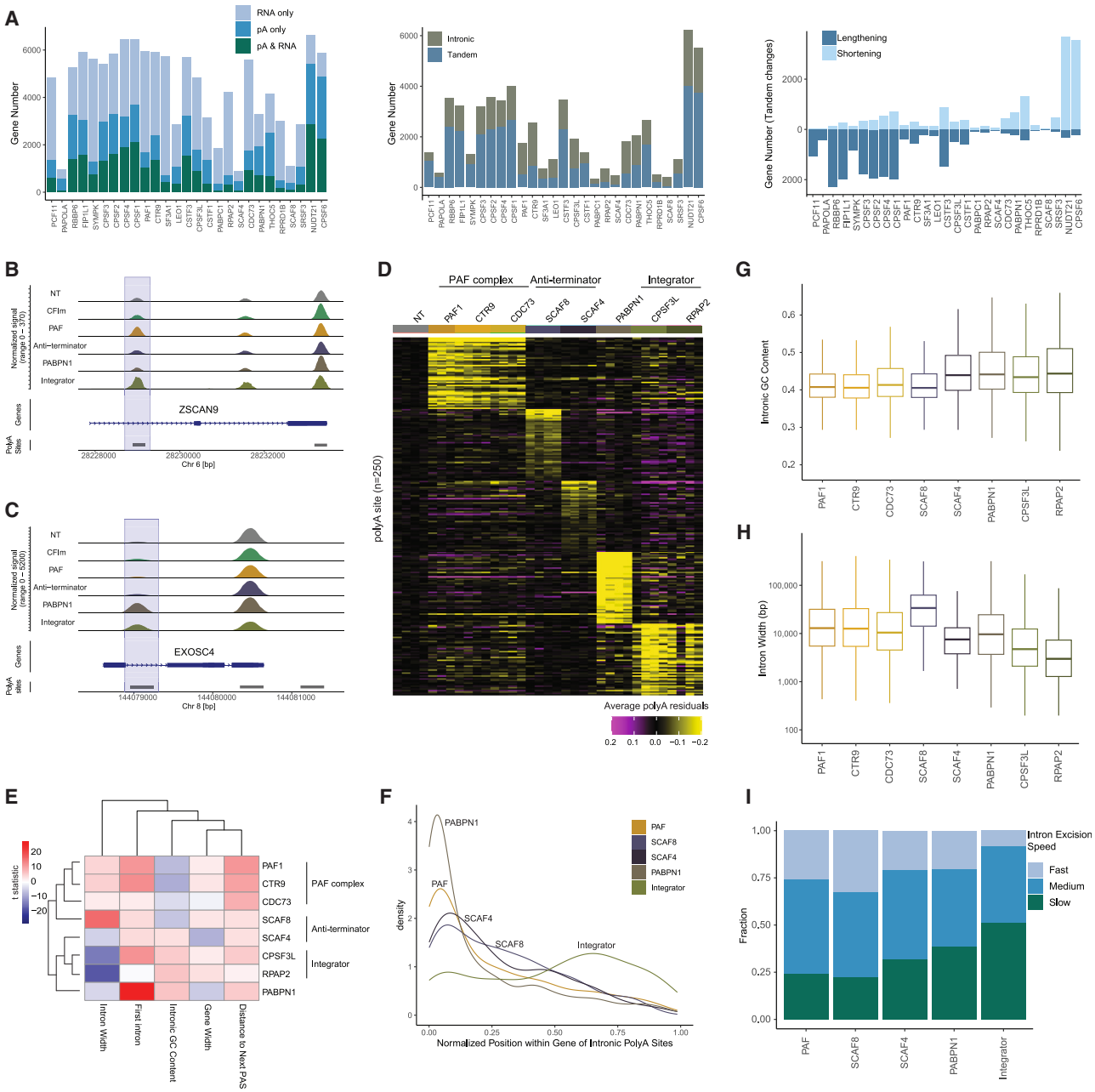


Figure 3. Characterizing tandem and intronic alternative polyadenylation in CPA-Perturb-seq

(A) (Left) Number of genes with changes in polyA site usage, gene expression, or both after perturbation of each regulator in the HEK293FT dataset. (Middle) Number of genes with perturbation-driven changes in intronic or tandem polyA site usage. (Right) Number of genes with changes in tandem polyA site usage, classified by 3′ UTR shortening or 3′ UTR lengthening.

(B and C) Read coverage plots showing differential usage of intronic sites (boxed) at the ZSCAN9 (B) and EXOSC4 (C) loci.

(D) Heatmap showing polyA residuals for intronic sites that are uniquely differentially utilized after perturbation of PAF complex members (PAF1/CTR9/CDC73), anti-terminators (SCAF4/SCAF8), PABPN1, and CPSF3L/RPAP2. Each heatmap cell shows the pseudobulk average of cells after grouping by sgRNA identity.

(E) Heatmap showing the importance of different features in predicting intronic polyA site usage for each perturbation. Color represents the t statistic for each covariate from a predictive linear model (STAR Methods), which was also used for hierarchical clustering.

(F) Metagene plot showing normalized position of intronic polyA sites with significant changes for each regulator.

(G) GC content for introns containing polyA sites with significant changes in usage for each regulator.

(H) Width for introns containing polyA sites with significant changes in usage for each regulator.

(I) Fraction of introns containing polyA sites with significant changes in usage for each regulator, grouped by intron excision speed, as classified by Mukherjee et al.[62]

See also Figure S4.

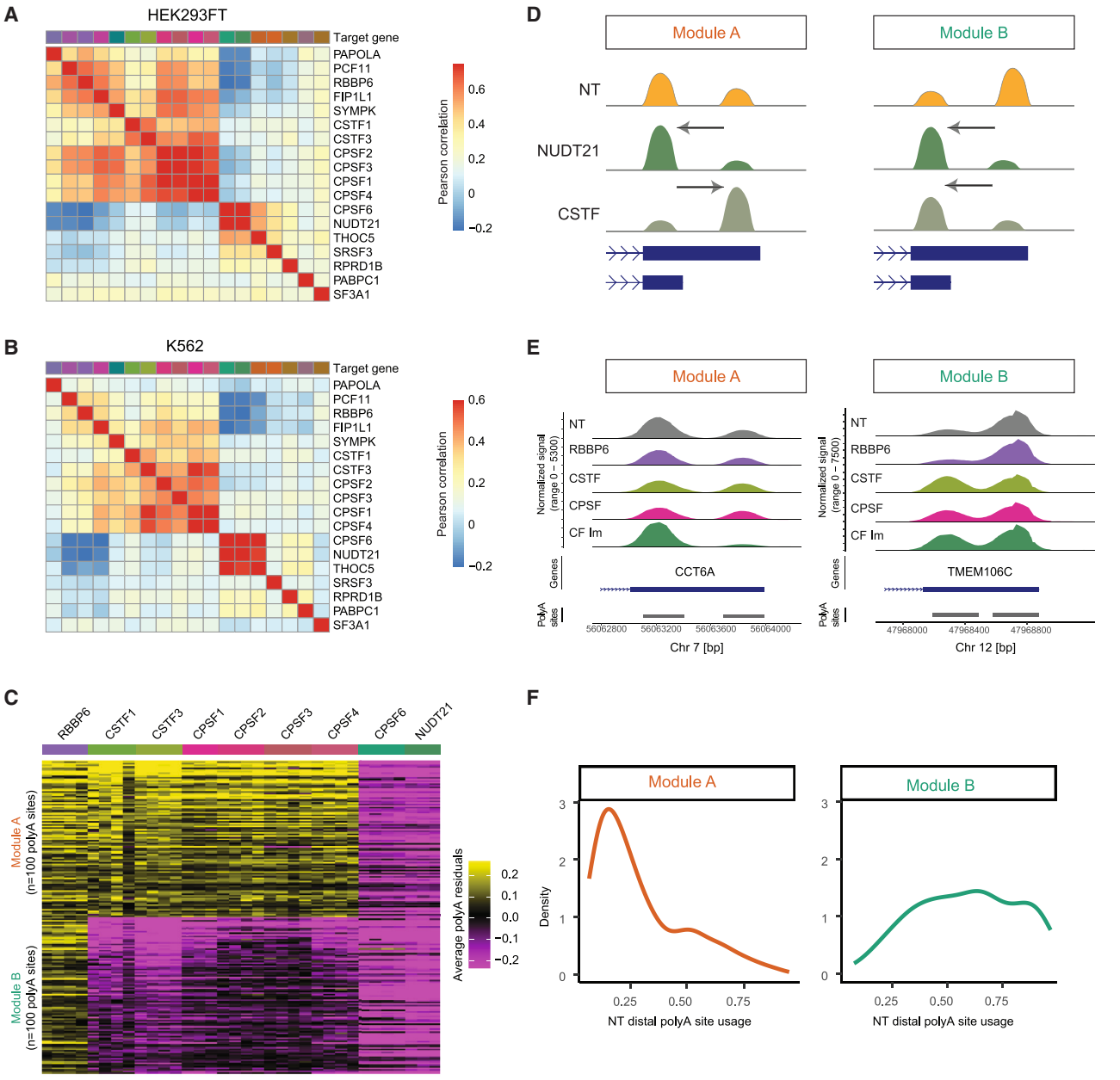


Figure 4. Modules of co-regulated polyA sites exhibit functional differences

(A) Pearson correlation matrix depicting the relationships between tandem perturbations in HEK293FT cells. Correlations are calculated using the linear model coefficients learned during differential polyadenylation analysis (STAR Methods). Genes are ordered via hierarchical clustering.

(B) Same as (A), but the correlation matrix is generated from K562 polyA residuals.

(C) Heatmap showing polyA residuals for distal peak sites in module A genes (CSTF and CPSF act in the opposite direction from CPSF6/NUDT21) and module B genes (CSTF and CPSF act in the same direction as CPSF6/NUDT21). Top 100 polyA sites (ranked by CSTF perturbation) are shown for each module.

(D) Schematic diagram of genes belonging to modules A and B.

(E) Read coverage plots showing polyA site usage of representative genes belonging to modules A (left, CCT6A) and B (right, TMEM106C).

(F) Density plot showing distal site usage in NT control cells for genes belonging to module A (left) vs. module B (right).

See also Figure S4.

We found that the correlation structure was driven not exclusively by global preferences toward shortening and lengthening but also by site-specific differences in perturbation response. Perturbation of RBBP6 (preference toward 3′ UTR lengthening) and CFIm complex members CPSF6 and NUDT21 (preference toward 3′ UTR shortening) showed strongly anticorrelated responses, reflecting globally opposing regulation. By contrast, CSTF and CPSF complex members (preference toward 3′ UTR

lengthening) showed only weak anticorrelation with CFIm members, reflecting more complex patterns of co-regulation.

To further explore this, we considered the set of genes that exhibited transcriptional shortening after CFIm perturbation (STAR Methods). We divided these genes based on CSTF perturbation response, and we observed an expected module (Figures 4C–4E, module A) of 323 polyA sites (20%) where CSTF perturbation resulted in an opposing lengthening response. However, we also identified a module (module B) of 149 (9%) genes where CSTF perturbation resulted in shortening, phenocopying CFIm perturbation. The remaining 71% of sites did not exhibit changes in utilization upon CSTF perturbation. We observed reproducible patterns at the same loci in K562 cells (Figure S4J).

Moreover, we found that genes that exhibit 3′ UTR lengthening upon CSTF perturbation (module A) strongly favored proximal site usage in NT cells, while genes with the opposite CSTF perturbation phenotype (module B) exhibited distal site bias (Figures 4F and S4K). Variation in the degree of distal bias at these genes is likely driven by differences in the relative strength of CSTF activity at their proximal and distal sites.

APARENT-Perturb reveals an interactive *cis*-regulatory code

Our identification of reproducible patterns of differential polyadenylation emphasizes the role of local sequence in determining a polyA site's responsiveness to perturbations. We sought to extend deep learning models that accurately predict genome-wide patterns of alternative polyadenylation in baseline conditions[41–46] to predict perturbation response. 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.[78]

To predict baseline polyA site usage in unperturbed cells, we used the APARENT2 model, a residual neural net originally trained on MPRA datasets measured in HEK293FT cells.[43] Inspired by the MTSplice model,[79] we then trained a new ensemble-based multi-task perturbation network (APARENT-Perturb) that predicts polyA site usage in our 10 strongest perturbations, using 200 nucleotide sequences aligned on the core hexamer and the APARENT2 baseline scores as input (Figure 5A). APARENT-Perturb accurately predicted the isoform proportion of polyA sites for held-out genes in the NT condition ($R_S = 0.70$) and in perturbations ($0.65 \leq R_S \leq 0.73$, as measured by 10-fold cross-validation (Figure 5B and S5A).

Next, we performed *in silico* mutagenesis (ISM), which yields a set of nucleotide-level "attribution scores," reflecting the contribution of each base to the model's prediction.[80,81] 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 TGTA motif that is predicted to drive responsiveness to NUDT21 perturbation and a distinct downstream GT-rich region motif that drives responsiveness to CSTF3 perturbation (Figures 5C and S5B). For each perturbation, we averaged ISM scores across loci to identify regions that harbored important sequence elements (Figures 5D and 5E). We next used a motif discovery tool, TF-MoDISco,[78,82] to cluster the attribution scores of each perturbation into a set of salient motifs (Figures 5D, S5C, and S5D). These results recapitulate and extend previously established binding motifs and positions.[2,4,46] For example, CSTF1 and CSTF3 displayed a peak of importance in the downstream region with T- or GT-rich sequences among their top motifs. NUDT21 and CPSF6 displayed high average importance in the upstream region of polyA sites but also extended the canonical TGTA motif with A- and T-rich flanks on both ends. Intriguingly, perturbation responses of NUDT21 and CPSF6 were also driven by sequence elements located approximately 30–50 bp downstream (downstream element [DSE]). This DSE overlaps with a region of predicted importance for CSTF perturbation. This reflects a co-enrichment of functional sequences for both complexes at the same sites (Figures S5E–S5G), suggesting a regulatory interaction between these factors.

We previously observed that CSTF and CFIm complex members could jointly regulate polyA sites in either the same or opposing directions (Figures 4C–4F). 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. 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, $p < 2.0 \times 10^{-5}$, Wilcoxon two-sided rank-sum test). By contrast, we observed increased attribution scores for module B genes at distal sites (Figure 5F, right, $p < 1.6 \times 10^{-4}$). Taken together, these findings suggest a model where the sequence content at proximal polyA sites is particularly important in establishing both proximal bias and perturbation response.

Finally, we simulated individual and pairwise motif insertions to identify epistatic interactions between CPA regulators, as has been done for transcription factors.[78,83] For example, the CFIm complex includes a NUDT21 homodimer, but it is unclear if and how multiple TGTA motifs affect binding.[84,85] APARENT-Perturb assigned higher importance scores to NUDT21 motifs with A- and T-rich flanks, so we tested multiple possible flanking sequences when performing insertions.

When inserting adjacent TGTA motifs at short distances, we observed synergistic effects on NUDT21 perturbation when both motifs were surrounded by GC-rich sequences, while an AT-rich context was associated with sub-additive interactions (Figures 5G, S5H, and S5I). Insertions of the canonical core hexamer and GT-rich DSE element, both of which have been individually associated with RBBP6 and CSTF regulation,[24] exhibited a cooperative epistatic relationship maximized at 20-bp insertion distance (Figures 5H and S5J). We verified each of these results using polynomial feature regression (Figures S5K–S5L). We conclude that application of deep learning models to Perturb-seq datasets can reveal a *cis*-regulatory landscape that encodes complex patterns of co-regulation across multiple complexes.

Validating sequence predictions by massively parallel screening with perturbations

Our results suggest that APARENT-Perturb identifies sequences driving polyA site choice, assigns these sequences to specific

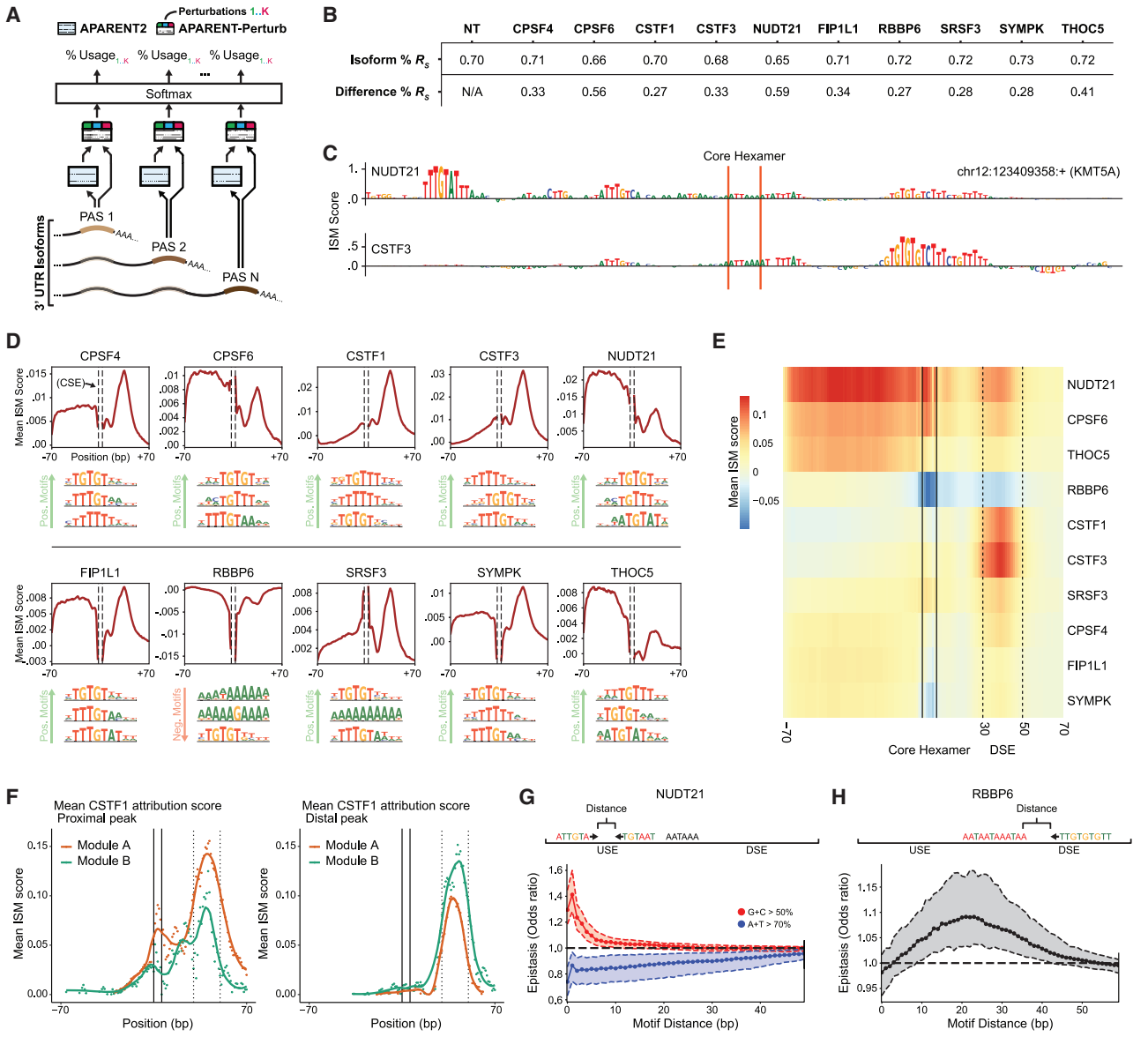


Figure 5. A multi-task neural network predicts perturbation responses from RNA sequence

(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 2 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 3 top MoDISco motifs are shown for each perturbation (STAR 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 (left) and distal (right) sites. Locations of the core hexamer and downstream sequence elements (DSEs) 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 TGTA motifs in either GC-rich (red) or AT-rich (blue) contexts. 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 (STAR Methods).

(H) Epistasis analysis of canonical hexamers and GT-rich motifs, based on the RBBP6 perturbation.

See also Figure S5.

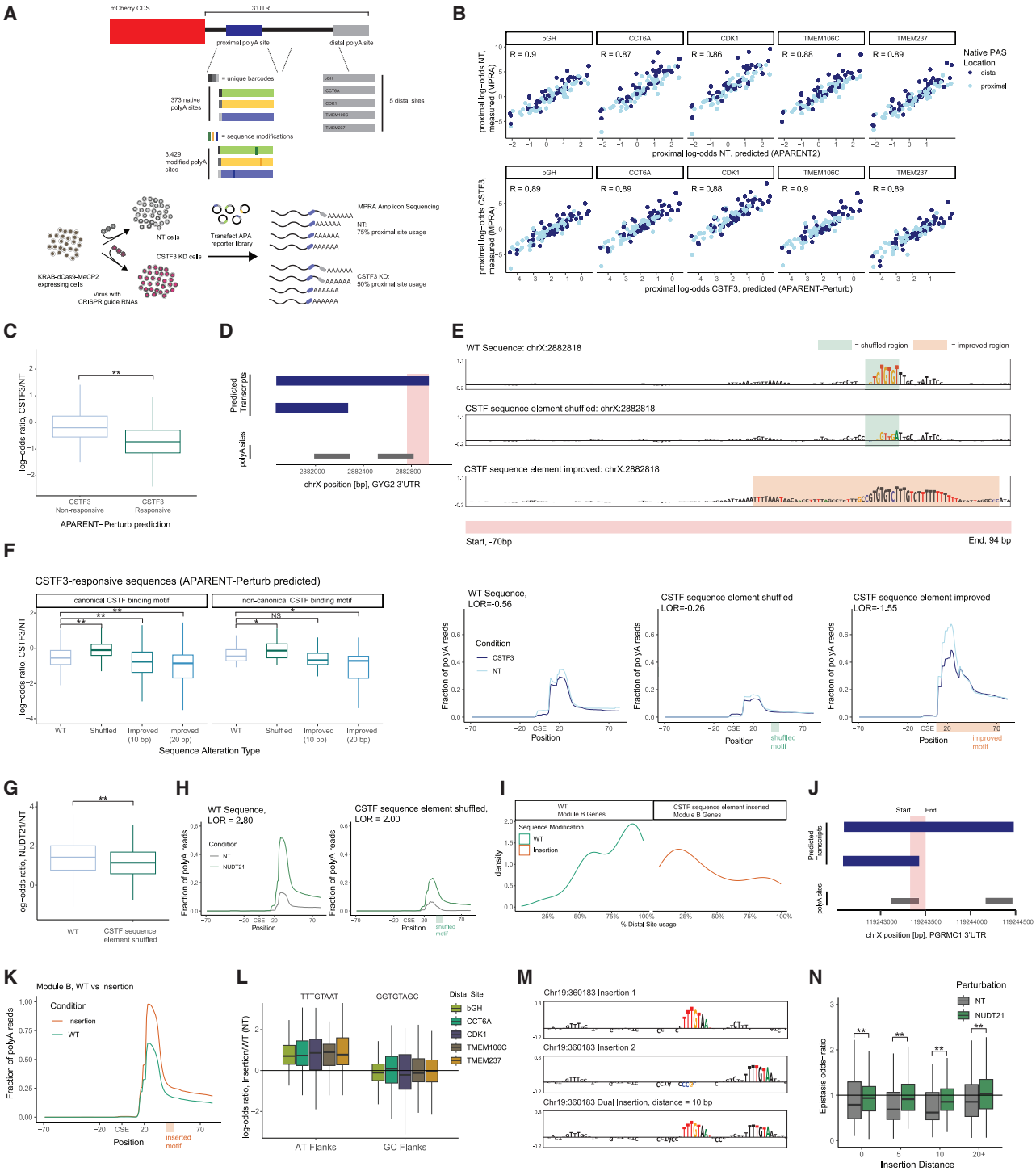


Figure 6. Validating APARENT-Perturb by performing an MPRA in multiple genetic contexts

(A) Schematic of massively parallel reporter assay (MPRA) used to validate APARENT-Perturb.

(B) Proximal polyA site usage (log odds) for 373 WT loci, as predicted by APARENT-Perturb (x axis) and measured by the MPRA (y axis). Predictions are accurate in both NT (top) and CSTF3-perturbed (bottom) samples and for both native proximal and distal sites.

(C) Change in proximal polyA site usage (log-odds ratio) comparing CSTF3 and NT samples for both sequences predicted by APARENT-Perturb to be responsive to CSTF3 perturbation (n = 107, right) vs. nonresponsive (n = 109, left). ** indicates p value <0.0001, Wilcoxon test comparing log-odds ratio of responsive to neutral sequences.

(D) Gene model of GYG2 3′ UTR, based on inferred cleavage sites from CPA-Perturb-seq. Region highlighted in red was inserted into the MPRA construct.

(legend continued on next page)

regulators, and identifies interactions between them. Validating these predictions at any locus requires two components: demonstrating that mutating sequence elements with high predicted importance alters polyA site usage and, additionally, showing that these alterations are dependent on the presence of the assigned regulator. We designed modified sequences at 373 loci (resulting in a total of 3,802 wild-type [WT] and mutated sequences) and leveraged a previously described reporter construct,[43] with minor modifications, to scalably validate APARENT-Perturb predictions (STAR Methods; Figures 6A and S6A). We performed the MPRA in samples with and without CRISPRi perturbation of CSTF3, enabling us to explore the effects of each sequence mutation with a genetic perturbation.

We inserted each of 3,802 test loci into the reporter's proximal site and used constructs with five distal sites of varying strength[42] (STAR Methods; Figure S6B; Table S5). For each genetic condition, we transfected cells with our MPRA library, performed two biological replicates, and observed reproducible polyA site quantifications (R = 0.96–0.98; Figure S7C).

First, we compared our APARENT-Perturb predictions of CSTF3 perturbation response to our MPRA. In both NT control cells (average R = 0.88) and CSTF3KD cells (average R = 0.89), we observed clear agreement with predicted values for each of the five distal sites (Figure 6B). Predicted responsive loci exhibited shifts after CSTF3 perturbation ($p < 3.4 \times 10^{-37}$), while predicted nonresponsive loci exhibited minimal changes (Figure 6C).

We next tested whether APARENT-Perturb could identify specific sequence elements driving this relationship. At each locus, we identified and scrambled the 9-bp sequence element with the maximal ISM score for the CSTF3 perturbation (STAR Methods; example in Figures 6D and 6E). These mutations correctly abrogated the difference between NT and CSTF3-perturbed cells (Figures 6F and S6D). Moreover, we also designed "super-responsive" loci, analogous to the model-guided design of synthetic enhancer sequences for TF binding.[86,87] We engineered 10 or 20 single-base-pair mutations to native loci that APARENT-Perturb predicted would increase the responsivity of the locus (STAR Methods). These sequences exhibited increased alterations to polyA site usage in the MPRA upon CSTF3 perturbation (Figure 6F).

We also found that APARENT-Perturb successfully identified CSTF3-responsive elements that lack the GT- or T-rich motif that has been linked to CSTF binding[88,89] (10% of responsive sequences, Figure 6F). Conversely, a subset of sequences that we predict and validate to be nonresponsive to CSTF3 perturbation contain a canonical DSE (Figure S6E). This highlights that APARENT-Perturb outperforms generic motif-based approaches to predict regulator activity.

Our previous analyses (Figures 5D, 5E, and S5D–S5F) suggested that sequence features driving CSTF regulation are also important for NUDT21 responsivity. Testing the same constructs in NUDT21 KD cells revealed that scrambling CSTF motifs also affected the responsivity to NUDT21 perturbation ($p < 3 \times 10^{-11}$, Figures 6G, 6H, and S6G). These regulatory elements were located downstream of the canonical hexamer and did not contain downstream TGTA motifs (Figure S6H). This effect was reduced compared to CSTF3 perturbation, possibly explained by remaining NUDT21 regulatory motifs (Figure S6F). This likely reflects a sequence-driven indirect regulatory interaction between NUDT21 and CSTF3.

We also tested our hypothesis that differential CSTF binding determines the proximal vs. distal bias we observed in module A vs. module B genes, respectively. To the proximal site of module B genes, we inserted a 9-bp sequence predicted by APARENT-Perturb to drive CSTF3 responsiveness. Strikingly, we found that this single sequence modification reversed the difference between the modules, shifting loci from distally biased (WT sequences) to proximally biased (mutant sequences) (Figures 6I–6K and S6I). The effect of these alterations was reduced in CSTF3-perturbed cells, highlighting that CSTF3 activity drives this behavior (Figure S6J). By contrast, shuffling

(E) (Top) Attribution scores from the APARENT-Perturb CSTF model at the distal site of the GYG2 gene (chrX:2882818), both for the wild-type (WT) sequence (top) and upon making sequence alterations. These included shuffling a predicted CSTF response element to minimize the effect of perturbation (middle) and designing synthetic mutations to maximize CSTF3 response. Colored nucleotides indicate nucleotides that were altered. (Bottom) Visualization of MPRA data at the GYG2 locus. Lines represent the fraction of reads that include a polyA sequence (polyA reads), which indicates proximal cleavage, for NT and CST3F-perturbed cells (STAR Methods).

(F) Effect of performing the sequence alterations described in (E) at 107 CSTF3-responsive sites, both with (left) and without (right) the canonical CSTF binding motif. The log-odds ratio of CSTF3 compared to NT (y axis) is shown for improved and shuffled sequence alterations (x axis). ** indicates p value <0.0001; * indicates p value = 0.0001–0.05; NS, not significant.

(G) Log-odds ratio comparing proximal polyA site usage in NUDT21 to NT samples (y axis), both for WT sequences and after shuffling CSTF sequence elements, as in (F). ** indicates p value <0.0001.

(H) Fraction of polyA reads, indicating proximal cleavage, for the FAM13C locus in both NUDT21-perturbed and NT cells, both for WT sequence and after shuffling CSTF sequence element.

(I) Density of distal polyA site usage for WT module A sequences (n = 47, green) and after inserting CSTF-responsive sequence elements into the proximal polyA site (right, orange).

(J) Gene model of the 3′ UTR of PGRM1 (a module B gene) based on inferred cleavage sites from CPA-Perturb-seq. Region highlighted in red was inserted into the MPRA construct.

(K) Effect of insertion of a CSTF-responsive sequence element into the construct depicted in (J) on the fraction of polyA reads (y axis).

(L) Effect of inserting NUDT21 motif (TGTA) into neutral sequences (y axis, log-odds ratio of polyA site usage for insertion in NT cells) when the motif was surrounded by AT-rich (left) and GC-rich (right) flanks, by distal site.

(M) APARENT-Perturb ISM scores for sequences with individual or dual insertions of the TTTGTAAT motif at the PIP5K1C locus.

(N) Epistasis odds ratio (y axis) of inserting TGTA motifs with AT-rich flanks at multiple distances (x axis) for both NT (gray) and NUDT21 (green) samples. ** indicates p value <0.0001 for one-sided t test of epistasis odds ratio = 1.

See also Figure S6.

the predicted CSTF response elements of module A proximal sites caused a shift toward distal polyA site usage (Figure S6I). This demonstrates that APARENT-Perturb successfully identified sequence elements that dictate proximal vs. distal bias in addition to perturbation response.

Lastly, we tested APARENT-Perturb's predictions of epistatic relationships between sequence motifs. We performed single and dual insertions of the NUDT21 TGTA motif into 49 loci, testing both AT-rich flanks and GC-rich flanks. Consistent with APARENT-Perturb's predictions, single insertions of the AT-rich motif resulted in significantly stronger ($p < 3.0 \times 10^{-54}$) alterations in polyA site usage compared to insertions of the GC-rich motif (Figure 6L). The strength of this insertion was weakened in NUDT21-perturbed cells compared to NT controls (Figure S6K). We also observed sub-additive effects when performing dual insertions of TGTA motifs with AT-rich flanks at short distances (Figures 6M and 6N), although insertion of GC-flanking motifs exhibited minimal effects.

Identification of CPA regulators from genome-wide screening datasets

To understand how additional regulators affect polyA site usage, we reanalyzed a recently published genome-wide Perturb-seq (GWPS) dataset[47] performed in K562 cells. We computed polyA residuals and used these as input to differential polyadenylation analysis (STAR Methods). As the GWPS dataset contained far fewer cells per perturbation (median 94 cells vs. 1,142 in our dataset for the 36 overlapping perturbations), we identified substantially fewer genes exhibiting changes in polyA site usage (median of 72 genes per overlapping perturbation compared to 1,389 in our data). However, the GWPS dataset still enabled accurate characterization of each regulator. For example, we observed concordant biases toward 3′ UTR shortening or lengthening induced by perturbation of tandem polyadenylation regulators across both datasets (Figure 7A).

To focus on regulators that directly modify RNA, we restricted our analysis to a set of 1,280 RNA binding proteins.[90] We identified 134 perturbations with polyA site usage changes in at least 100 genes (Figure 7B; Table S4), including highly correlated perturbations (Figures 7C and S7A). Perturbation of CFIm complex members CPSF6 and NUDT21 and transcription-export complex member THOC3 exhibited correlated 3′ UTR shortening (Figures S7B and S7D), while a module consisting of up-frameshift complex members, small ribosomal subunits, and the ribosome maturation factor were associated with 3′ UTR lengthening (module 4; Figure S7G). While these genes are well-studied regulators of translational control and RNA stability, none have been previously associated with regulating polyA site choice. We also identified a module with components of the large ribosomal subunit along with the translation initiation factor EIF6 (module 2; Figure S7E), suggesting tight crosstalk between alternative polyadenylation and multiple RNA regulatory processes.

We found correlated perturbation responses between members of the poly(A) tail exosome targeting (PAXT) complex (module 3), which also contained nuclear cap-binding complex member NCBP2 and splicing regulator MBNL1.[68,91–93] Perturbation of this module was associated primarily with upregulation of intronically polyadenylated transcripts (Figures 7B, 7D, and S7C). This response is likely driven by the PAXT complex's role in degrading prematurely terminated RNA transcripts,[68,94] although the surveillance machinery that specifically distinguishes premature transcripts remains unknown.[95]

Lastly, we aimed to investigate our previous hypothesis that intronic polyA site changes associated with integrator perturbations were due to changes in splicing dynamics (Figures 3D and 3I). Our CPA-Perturb-seq integrator signature was most correlated with members of the survival motor neuron (SMN) complex and other integrator complex members, which consisted of an additional module (module 5; Figures 7E, 7F, and S7H). We confirmed that the set of splice-module-regulated sites in GWPS uniquely overlapped with our integrator perturbation response signature (Figure 7G). Finally, we observed that these sites were located in introns with weaker canonical 5′ donor splice site scores, consistent with reduced splicing efficiency (Figure 7H).[96,97] This provides orthogonal support for our hypothesis that integrator perturbation changes polyA site usage within introns that are inefficiently spliced, demonstrating their sensitivity to changes in splicing dynamics.

We conclude that 3′ scRNA-seq data can be combined with tailored computational pipelines to explore cellular heterogeneity in polyA site usage, and we have developed an open-source R package, PolyA Site analysis using relative Transcript Abundance (PASTA), that implements the analytical methods described in this manuscript. PASTA is fully compatible with our analytical toolkit Seurat,[98] and the software release includes a vignette demonstrating how users can explore cellular heterogeneity in alternative polyadenylation in a dataset of circulating human peripheral blood mononuclear cells (STAR Methods). 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 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. We introduce a statistical framework to quantify changes in relative polyA site usage across regulators at single-cell resolution and identify modules of co-regulated polyA sites.

Our CPA-Perturb-seq dataset revealed striking heterogeneity in perturbation response, including the number, type, and directionality of changes associated with each regulator. This demonstrates that alternative polyadenylation is not uniformly regulated, where all polyA sites are equally sensitive to perturbation of core regulators. Instead, we consistently observed evidence of distinct regulatory responses across modules of polyA sites.

Using our deep neural network, APARENT-Perturb, we found that this local regulatory structure is encoded in part by sequence-specific elements surrounding the cleavage site. By integrating previously trained sequence-based models with our perturbation data, we directly learn associations between sequence elements and regulators, providing a more mechanistic understanding of *cis*-regulatory element function,

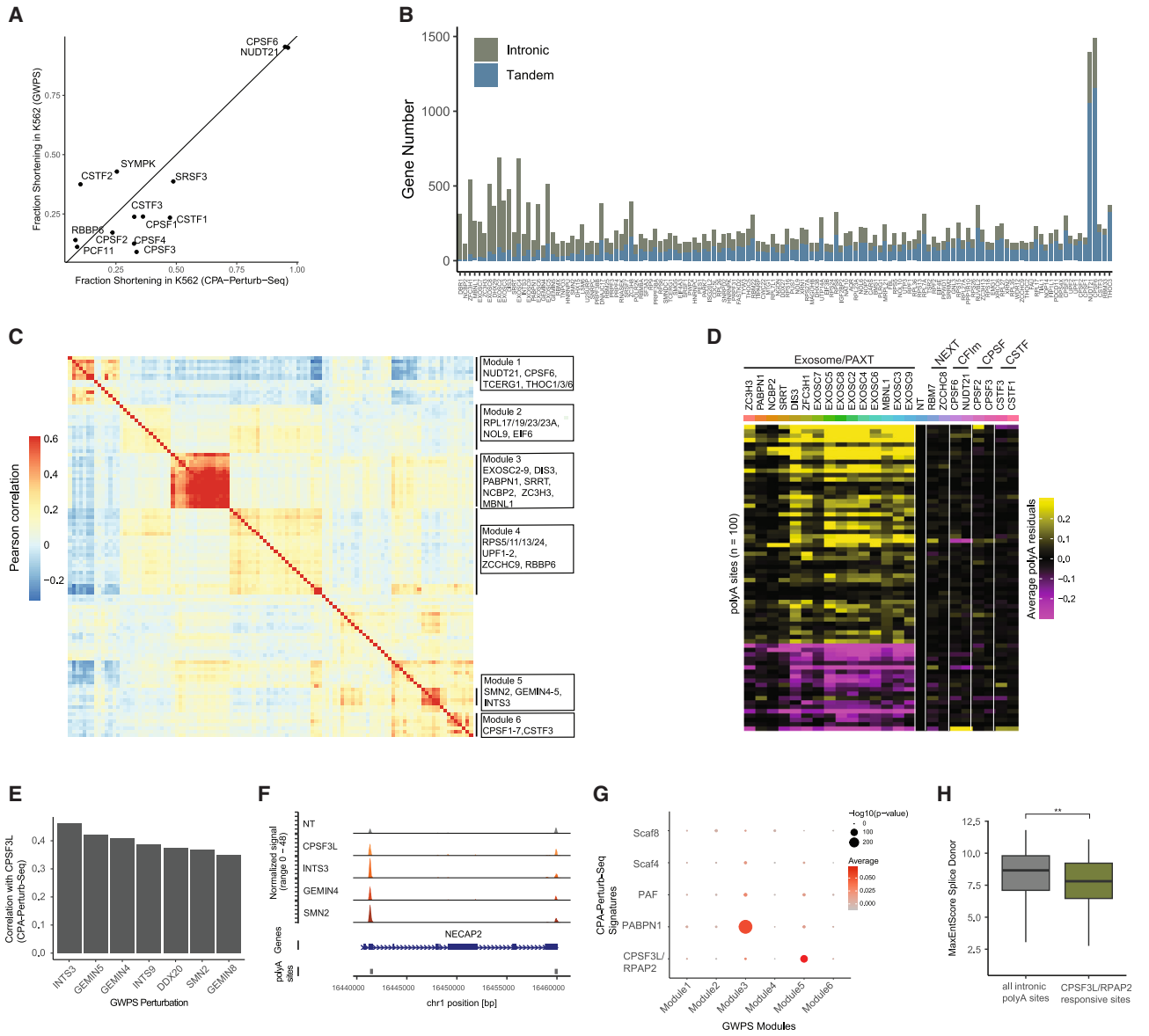


Figure 7. Characterizing heterogeneity in relative polyA site usage in genome-scale Perturb-seq datasets

(A) 3′ UTR shortening preference observed after perturbing tandem regulators in the CPA-Perturb-seq dataset (x axis) and the GWPS dataset (y axis).

(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 Figure S7A.

(D) Representative polyA sites (n = 100) whose usage increased upon perturbation of exosome/PAXT complex members.

(E) CPSF3L perturbation (from CPA-Perturb-seq) was most correlated (y axis) with other members of the integrator complex and spliceosome factors in GWPS data.

(F) Representative read coverage plot depicting shared changes in polyA site usage after perturbation of CPSF3L, other integrator complex members, and SMN complex members.

(G) Enrichment of CPA-Perturb-seq intronic signatures (y axis) in GWPS module-responsive sites. Dot size corresponds to −log10(p value) from hypergeometric enrichment test, and color corresponds to the average polyA residuals for each signature.

(H) MaxEnt scores for splice donor sites (5′) of all intronic polyA sites quantified in our dataset vs. those with significantly increased usage in CPSF3L/RPAP2 perturbation. ** indicates p value <0.0001, Wilcoxon test.

See also Figure S7.

including interactions between regulators. We note that this strategy could be extended to additional sequence-based deep learning models.[99,100]

While our analyses aimed to focus on regulatory mechanisms that influenced CPA decisions, we repeatedly observed cases where additional RNA regulatory processes altered the relative abundance of alternatively polyadenylated transcripts. We found that perturbation of proteins with roles in RNA polymerase elongation, RNA export, translation, and splicing resulted in differential usage of polyA sites, highlighting extensive interdependencies between RNA regulatory processes. Future work may exploit these interdependencies to infer RNA kinetic parameters from 3′ scRNA-seq data. More broadly, our statistical method may be extended to characterize additional sources of transcriptomic diversity, such as alternative splicing.

Our approach has limitations for measuring and interpreting changes in polyadenylation. Although correlated perturbation responses across regulators likely reflect shared function, we cannot exclude the possibility that indirect effects (where the perturbation of one regulator affects the expression of another) also can affect correlation structure. Perturbations can introduce changes to both production and degradation rates of individual transcripts, and Perturb-seq cannot conclusively distinguish between these two phenomena. Also, perturbations that substantially shorten the length of the polyA tail may generate biases in transcript capture that our study cannot address. Lastly, our annotation of polyA sites is based on the polyA_DB database, and we may occasionally misclassify tandem polyA sites as intronic. Future studies that combine the CPA-Perturb-seq workflow with RNA metabolic labeling[101,102] or long-read sequencing[103] can more accurately quantify isoforms and represent exciting extensions of our work.

Looking forward, we believe that scRNA-seq analyses of post-transcriptional regulation from perturbation screens and primary samples will be mutually informative. Functional genomics tools like Perturb-seq are well suited to identify targets of molecular regulators. 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, such as disease conditions. Integration of these datasets therefore represents a potential path forward for systematic reconstruction of regulatory networks guiding the RNA life cycle.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.06.005.

ACKNOWLEDGMENTS

The authors would like to acknowledge Torben Heick Jensen, Robert Bradley, Christina Leslie, and Christine Mayr for thoughtful discussions related to this work. 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 grant 2U24HG007234.

AUTHOR CONTRIBUTIONS

H.-H.W. and R.S. conceived the study. M.H.K., H.H.W., J.L., S.K., A.H., and Y.H. performed computational work, supervised by A.K. and R.S. H.-H.W., C.D., and I.M. performed experimental work, supervised by R.S. All authors participated in interpretation and writing and editing the manuscript.

DECLARATION OF INTERESTS

In the past 3 years, R.S. has received compensation from Bristol-Myers Squibb, ImmunAI, Resolve Biosciences, Nanostring, 10x Genomics, Neptune Bio, and the NYC Pandemic Response Lab. R.S., H.-H.W., and Y.H. are co-founders and equity holders of Neptune Bio. A.K. is a scientific cofounder of Ravel Biotechnology; is on the scientific advisory board of PatchBio, SerImmune, AINovo, TensorBio, and OpenTargets; is a consultant with Illumina; and owns shares in DeepGenomics, Immunai, and Freenome. J.L. is an employee of Calico Life Sciences, LLC, as of 11/21/2022. H.-H.W. and Y.H. are employees of Neptune Bio as of 8/1/2023.

Received: February 9, 2023
Revised: March 12, 2024
Accepted: June 5, 2024
Published: June 25, 2024

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. 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.
19. 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.
20. 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.
21. 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*, 629–631.
22. 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.
23. 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.
24. Di Giammartino, D.C., Li, W., Ogami, K., Yashinskie, J.J., Hoque, M., Tian, B., Manley, J.L., and Manley, J.L. (2014). RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3′ UTRs. Genes Dev. *28*, 2248–2260.
25. Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L., and Tian, B. (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.
26. 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.
27. Wang, R., Zheng, D., Wei, L., Ding, Q., and Tian, B. (2019). Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. Cell Rep. *26*, 2766–2778.e6.
28. Dubbury, S.J., Boutz, P.L., and Sharp, P.A. (2018). CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. Nature *564*, 141–145.
29. Takagaki, Y., Seipelt, R.L., Peterson, M.L., and Manley, J.L. (1996). The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell *87*, 941–952.
30. Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–668.
31. de Prisco, N., Ford, C., Elrod, N.D., Lee, W., Tang, L.C., Huang, K.-L., Lin, A., Ji, P., Jonnakuti, V.S., Boyle, L., et al. (2023). Alternative polyadenylation alters protein dosage by switching between intronic and 3′UTR sites. Sci. Adv. *9*, eade4814.
32. 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.
33. 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.
34. 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.
35. 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.
36. Wessels, H.-H., Méndez-Mancilla, A., Hao, Y., Papalexi, E., Mauck, W.M., III, Lu, L., Morris, J.A., Mimitou, E.P., Smibert, P., Sanjana, N.E., and Satija, R. (2023). Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. Nat. Methods *20*, 86–94.
37. 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.
38. 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.
39. Fansler, M.M., Zhen, G., and Mayr, C. (2021). Quantification of alternative 3′UTR isoforms from single cell RNA-seq data with scUTRquant. Preprint at bioRxiv. https://doi.org/10.1101/2021.11.22.469635.
40. Harrison, P., Williams, S., Powell, D., Albrecht, D., and Beilharz, T. (2019). Tools for identifying and characterizing alternative polyadenylation in

scRNA-Seq. F1000Res. *8*, 1142. https://doi.org/10.7490/f1000research.1117076.1.

41. Leung, M.K.K., Delong, A., and Frey, B.J. (2018). Inference of the human polyadenylation code. Bioinformatics *34*, 2889–2898.
42. 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.
43. 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.
44. Arefeen, A., Xiao, X., and Jiang, T. (2019). DeepPASTA: deep neural network based polyadenylation site analysis. Bioinformatics *35*, 4577–4585.
45. Li, Z., Li, Y., Zhang, B., Li, Y., Long, Y., Zhou, J., Zou, X., Zhang, M., Hu, Y., Chen, W., and Gao, X. (2022). DeeReCT-APA: Prediction of Alternative Polyadenylation Site Usage Through Deep Learning. Dev. Reprod. Biol. *20*, 483–495.
46. Vainberg Slutskin, I., Weinberger, A., and Segal, E. (2019). Sequence determinants of polyadenylation-mediated regulation. Genome Res. *29*, 1635–1647.
47. 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.
48. Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck, W.M., III, Wessels, H.-H., Hao, Y., Yeung, B.Z., Smibert, P., and Satija, R. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nat. Genet. *53*, 322–331.
49. Frohman, M.A., Dush, M.K., and Martin, G.R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA *85*, 8998–9002.
50. Scheer, H., De Almeida, C., Sikorska, N., Koechler, S., Gagliardi, D., and Zuber, H. (2020). High-Resolution Mapping of 3′ Extremities of RNA Exosome Substrates by 3′ RACE-Seq. Methods Mol. Biol. *2062*, 147–167.
51. Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. Cell Rep. *1*, 753–763.
52. Hwang, H.-W., Park, C.Y., Goodarzi, H., Fak, J.J., Mele, A., Moore, M.J., Saito, Y., and Darnell, R.B. (2016). PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. Cell Rep. *15*, 423–435.
53. 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.
54. Lause, J., Berens, P., and Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. *22*, 258.
55. 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.
56. Choudhary, S., and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. *23*, 27.
57. Mosimann, J.E. (1962). On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions. Biometrika *49*, 65–82.
58. 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.
59. 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.
60. Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. USA *99*, 12795–12800.
61. 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.
62. Mukherjee, N., Calviello, L., Hirsekorn, A., de Pretis, S., Pelizzola, M., and Ohler, U. (2017). Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nat. Struct. Mol. Biol. *24*, 86–96.
63. O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front. Endocrinol. *9*, 402.
64. Chen, C.Y., and Shyu, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem. Sci. *20*, 465–470.
65. 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.
66. 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. USA *116*, 14583–14592.
67. Baillat, D., Hakimi, M.-A., Näär, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265–276.
68. Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J.S., Sandelin, A., and Jensen, T.H. (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol. Cell *64*, 520–533.
69. Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. (2012). Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. Mol. Cell *45*, 111–122.
70. Wang, X., Qi, Y., Wang, Z., Wang, L., Song, A., Tao, B., Li, J., Zhao, D., Zhang, H., Jin, Q., et al. (2022). RPAP2 regulates a transcription initiation checkpoint by inhibiting assembly of pre-initiation complex. Cell Rep. *39*, 110732.
71. Elrod, N.D., Henriques, T., Huang, K.-L., Tatomer, D.C., Wilusz, J.E., Wagner, E.J., and Adelman, K. (2019). The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. Mol. Cell *76*, 738–752.e7.
72. Stein, C.B., Field, A.R., Mimoso, C.A., Zhao, C., Huang, K.-L., Wagner, E.J., and Adelman, K. (2022). Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci. Mol. Cell *82*, 4232–4245.e11.
73. Jeronimo, C., Forget, D., Bouchard, A., Li, Q., Chua, G., Poitras, C., Thérien, C., Bergeron, D., Bourassa, S., Greenblatt, J., et al. (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. Mol. Cell *27*, 262–274.
74. de Klerk, E., Venema, A., Anvar, S.Y., Goeman, J.J., Hu, O., Trollet, C., Dickson, G., den Dunnen, J.T., van der Maarel, S.M., Raz, V., and 't Hoen, P.A.C. (2012). Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Res. *40*, 9089–9101.
75. Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kühn, U., Menzies, F.M., Oude Vrielink, J.A.F., Bos, A.J., Drost, J., Rooijers, K., et al. (2012). The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell *149*, 538–553.
76. 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.

77. Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. *1*, 543–556.

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

79. 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.

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

81. 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.

82. Shrikumar, A., Tian, K., Shcherbina, A., Avsec, Ž., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. (2018). TF-MoDISco v0.4.4.2-alpha: Technical Note. Preprint at arXiv.

83. de Almeida, B.P., Reiter, F., Pagani, M., and Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat. Genet. *54*, 613–624.

84. 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. USA *107*, 10062–10067.

85. 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.

86. de Almeida, B.P., Schaub, C., Pagani, M., Secchia, S., Furlong, E.E.M., and Stark, A. (2024). Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. Nature *626*, 207–211.

87. Taskiran, I.I., Spanier, K.I., Dickmänken, H., Kempynck, N., Pančíková, A., Ekşi, E.C., Hulselmans, G., Ismail, J.N., Theunis, K., Vandepoel, R., et al. (2024). Cell-type-directed design of synthetic enhancers. Nature *626*, 212–220.

88. MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol. Cell Biol. *14*, 6647–6654.

89. Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L., and Hovorun, D.M. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res. *31*, 1375–1386.

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

91. 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.

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

93. 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.

94. Ogami, K., Richard, P., Chen, Y., Hoque, M., Li, W., Moresco, J.J., Yates, J.R., III, 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.

95. 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.

96. Yeo, G., and Burge, C.B. (2003). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In RECOMB '03: Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (Association for Computing Machinery), pp. 322–331.

97. Eng, L., Coutinho, G., Nahas, S., Yeo, G., Tanouye, R., Babaei, M., Dörk, T., Burge, C., and Gatti, R.A. (2004). Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. Hum. Mutat. *23*, 67–76.

98. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., and Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. *42*, 293–304.

99. Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D.R. (2023). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.30.555582.

100. Celaj, A., Gao, A.J., Lau, T.T.Y., Holgersen, E.M., Lo, A., Lodaya, V., Cole, C.B., Denroche, R.E., Spickett, C., Wagih, O., et al. (2023). An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. Preprint at bioRxiv. https://doi.org/10.1101/2023.09.20.558508.

101. Herzog, V.A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T.R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S.L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. Nat. Methods *14*, 1198–1204.

102. Cao, J., Zhou, W., Steemers, F., Trapnell, C., and Shendure, J. (2020). Sci-fate characterizes the dynamics of gene expression in single cells. Nat. Biotechnol. *38*, 980–988.

103. Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., Sloan, S.A., et al. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat. Biotechnol. *36*, 1197–1202. https://doi.org/10.1038/nbt.4259.

104. Morris, J.A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D.A., Hao, S., et al. (2023). Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. Science *380*, eadh7699.

105. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341.

106. Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., and Doench, J.G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat. Commun. *9*, 5416.

107. Jost, M., Santos, D.A., Saunders, R.A., Horlbeck, M.A., Hawkins, J.S., Scaria, S.M., Norman, T.M., Hussmann, J.A., Liem, C.R., Gross, C.A., and Weissman, J.S. (2020). Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. Nat. Biotechnol. *38*, 355–364.

108. Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., III, Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*, 224.

109. McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al.

(2019). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626.

110. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

111. Wang, R., Zheng, D., Yehia, G., and Tian, B. (2018). A compendium of conserved cleavage and polyadenylation events in mammalian genes. Genome Res. *28*, 1427–1441.

112. Bronner, I.F., and Quail, M.A. (2019). Best Practices for Illumina Library Preparation. Curr. Protoc. Hum. Genet. *102*, e86.

113. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12.

114. Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR. Curr. Protoc. Bioinformatics *51*, 11.14.1–11.14.19.

115. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

116. Kim, J., Zhang, Y., Day, J., and Zhou, H. (2018). MGLM: An R Package for Multivariate Categorical Data Analysis. R J. *10*, 73–90.

117. Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017). Regression Models For Multivariate Count Data. J. Comput. Graph Stat. *26*, 1–13.

118. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

119. Ptok, J., Theiss, S., and Schaal, H. (2020). VarCon: An R Package for Retrieving Neighboring Nucleotides of an SNV. Cancer Inform. *19*, 1176935120976399.

120. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Endura Electrocompetent Cells	Lucigen	Cat# 60242-2
One Shot Stbl3 Chemically Competent *E. coli*	ThermoFisher	Cat# C737303
Critical commercial assays		
Chromium Single-Cell 3′ v3 with Feature Barcoding	10X Genomics	PN-1000075, PN-1000153, PN-1000079
Deposited data		
HEK293FT- Perturb-seq	This paper	GEO: GSE269600
K562- Perturb-seq	This paper	GEO: GSE269600
HEK293FT- APARENT-Perturb MPRA	This paper	GEO: GSE269600
HEK293FT- 3′RACE	This paper	GEO: GSE269600
polyA_DB version 3.2	Wang et al.[7]	https://exon.apps.wistar.org/polya_db/v3/
HEK293 cells- Aseq	Martin et al.[51]	GEO: GSE37401
HEK293 cells- PAPERCLIP	Hwang et al.[52]	GEO: GSE66092
HEK293 cells- intron excision speeds	Mukherjee et al.[62]	GEO: GSE84722
K562 cells- Genome-scale Perturb-seq	Replogle et al.[47]	SRA: SRP376262
Experimental models: Cell lines		
K562 KRAB-dCas9-MeCP2	This paper	N/A
HEK293FT KRAB-dCas9-MeCP2	Wessels et al.[36]	N/A
Oligonucleotides		
See Table S3 for 3′ RACE primers	This paper	N/A
See Table S5 for MPRA primers	This paper	N/A
Recombinant DNA		
lentiGuideFB-Puro-A	Wessels et al.[36]	Addgene #192506
psPAX2	N/A	Addgene #12260
pMD2.G	N/A	Addgene #12259
APA Reference Library	Linder et al.[43]	Addgene #193784
lentiCRISPRi(v2)-Blast	Morris et al.[104]	Addgene #170068
Software and algorithms		
Cellranger v6.0.0	10X Genomics	https://www.10xgenomics.com/software
polyApipe v1.0	N/A	https://github.com/MonashBioinformaticsPlatform/polyApipe
CITE-seq-count v1.4.2	N/A	https://github.com/Hoohm/CITE-seq-Count
Sinto v0.9.0	N/A	https://github.com/timoast/sinto
Seurat v5.0	Hao et al.[98]	https://github.com/satijalab/seurat
Signac v1.12.0	Stuart et al.[105]	https://github.com/stuart-lab/signac
PASTA	This paper	https://github.com/satijalab/PASTA

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Rahul Satija (rsatija@nygenome.org).

Materials availability

No unique reagents were generated for this study.

Data and code availability

- The CPA-Perturb-seq datasets generated for this study are available for download at https://zenodo.org/record/7619593#.Y-P7Zi1h2X0. All raw data are avilable in the GEO database under the accession number GEO: GSE269600.
- CPA-Perturb-seq data can be explored via custom UCSC GenomeBrowser tracks, available at https://satijalab.org/cpa-perturb-seq.
- Seurat and PASTA are both available as open-source R packages at https://github.com/satijalab/seurat and https://github.com/satijalab/PASTA.
- Code to train and interpret the APARENT-Perturb model is available at https://github.com/johli/aparent-perturb.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell culture and maintenance

HEK293FT (HEK) and K562 cells were acquired from ThermoFisher (R70007) and ATCC (CCL-243), respectively. Monoclonal HEK293FT and K562 lines that constitutively express KRAB-dCas9-MeCP2 (CRISPRi) were generated as previously described.[36,104] HEK293FT and K562 cells were maintained in DMEM (Caisson DML23) or RPMI (Thermo Fisher 11875119), respectively, supplemented with 10% fetal bovine serum (Serum Plus II Sigma-Aldrich 14009C) and with no antibiotics at 37°C and 5% CO_2. Constitutive CRISPR Cas effector expressing cells were maintained using 5μg/mL Blasticidin S (ThermoFisher A1113903). Guide RNA expressing cells were grown using 1μg/mL puromycin (ThermoFisher A1113803).

METHOD DETAILS

sgRNA design, virus production, and transduction

We selected the most effective sgRNAs from previous genome-wide CRISPR-interference screens[106,107] along with non-targeting (NT) sgRNAs that showed no depletion. Individual sgRNAs were cloned into lentiGuideFB-Puro-A (Addgene #192506) as previously described.[36] All constructs were confirmed by Sanger sequencing. Individual plasmids were pooled accounting for the relative expected essentiality of target genes to compensate for the expected loss of cells with essential gene sgRNAs.

For pooled virus production we seeded 1×10^7 HEK293FT cells per 10 cm dish 12–18 h before transfection [60μL PEI, 6.4 μg psPAX2 (Addgene #12260), 4.4 μg pMD2.G (Addgene #12259) and 9.2 μg of the plasmid pool]. Six to 8 h post-transfection, the medium was exchanged for 10 mL of DMEM +10% FBS containing 1% bovine serum albumin (BSA). Viral supernatants were collected after additional 48 h, spun down to remove cellular debris for 5 min at 4°C and 1000 × g and passed through a 0.45 μm filter prior to storage at −80°C.

For single-cell experiments we transduced 3×10^6 HEK293FT or K562 CRISPRi cells per 12-well with an MOI of <0.1 (<10% survival) ensuring high coverage (cells per sgRNA) of at least 1000× and a single integration probability >95%. The cells were selected with 1μg/ml puromycin starting at 24 h post-transduction for at least 48 h for HEK293FT cells and at least 5 days for K562 cells. Cells were maintained with blasticidin and puromycin until the single-cell experiment 6–7 post transduction.

Direct capture Perturb-seq and sequencing

We performed scRNA-seq (10x Genomics Chromium Single Cell 3′ Gene Expression v3 with Feature Barcoding technology for CRISPR screening) 6–7 days post-transduction. Before the run, cell viability was determined (>95%). We leveraged Cell Hashing in order to super-load (∼40,000 cells/lane) the Chromium instrument.[108] Gene expression and sgRNA feature libraries were constructed following the manufacturer's protocol. Cell Hashing libraries (Hashtag-derived oligos, HTOs) libraries were prepared following the recommended protocol on cite-seq.com. All libraries were sequenced on Illumina NovaSeq S4 flow cells.

Processing of single-cell gene expression data

HEK293FT and K562 scRNA-seq and Feature Barcoding data was processed using CellRanger (v6.0.0, ''cellranger count''). HTO data was quantified using the *CITE-seq-count* package (v1.4.2). Count matrices were then used as input to the PASTA R package, which leverages both Seurat (v5.0)[98] and Signac (v1.12.0)[105] to perform downstream analyses. HTO and sgRNA counts were normalized using the centered log-ratio transformation approach. We retained cells with unique HTO and sgRNA assignments generated using the *MULTIseqDemux*[109] function in Seurat.

Quantifying polyA site usage

CellRanger output BAM files were used as input to polyApipe (https://github.com/MonashBioinformaticsPlatform/polyApipe), a computational workflow which quantifies polyA site usage at single-cell resolution. Briefly, this workflow comprises two steps.

First, polyApipe attempts to discover a set of polyA sites that are utilized in the dataset. All reads in the overall sample that contain a stretch of at least 5 softclipped A's at the 3′ end of the read that were not present in the genomic sequence are annotated as ''polyA reads.'' This subset of reads is used for peak calling. We required at least 5 ''polyA reads'' across all cells to form a polyA site. In cases

where two independently discovered sites fell within the defined counting region (here: 300 nucleotides upstream of the identified polyA site) the peak with the higher polyA read count was retained.

Second, polyApipe quantifies the read coverage, at each of these polyA quantification windows, for each single cell in the dataset. For counting, all reads were used, regardless of whether they were originally annotated as "polyA reads." The outputs of polyApipe are a tabular count file and counts matrix quantifying the usage of each polyA site at single-cell resolution, as well as a GFF file that describes the location of the inferred polyA sites. These outputs can be read via the PASTA package using the function *ReadPolyAPipe*.

To further ensure that our downstream analysis was conducted on *bona fide* polyA sites, we intersected these sites with those in PolyA_DB version 3.2 (polyAdbv3),[110,111] a catalog of sites identified from deep sequencing data using the (3′READS) method. We lifted over these sites from hg19 to hg38. We retained all polyA sites identified by polyApipe that were located within 50 nucleotides of a polyAdbv3 polyA site. We additionally used polyAdbv3 to assign a gene annotation to each polyA site, and further restricted our downstream analyses to polyA sites that are either located within an intron, or the last exon, of a transcript. The same set of polyA sites was used for analysis in K562 cells by providing the HEK293FT GFF file as an input to polyApipe.

In order to visualize the usage of polyA sites, we first used umitools to deduplicate the aligned BAM files. We then used the blocks function in Sinto (https://github.com/timoast/sinto) to construct a fragment file from the deduplicated BAM file. We use the function *PolyACoveragePlot* in PASTA to visualize read coverage throughout this manuscript (i.e., Figure 1C). In the bottom of these plots, we visualize the locations that represent the polyA sites identified by polyApipe.

3′ rapid amplification of cDNA ends (3′ RACE)

We performed 3′ RACE with an Illumina sequencing readout to validate the CPA-Perturb-seq workflow for identifying and quantifying polyA site usage. We pre-selected 10 genomic loci for targeted profiling, and in each case, designed a locus-specific primer that binds 100–150bp upstream of the predicted proximal cleavage site. As previously described, CRISPRi-expressing HEK293FT cells were transduced with lentivirus for guides targeting NUDT21 or NT (single guide per condition, 2 transduction replicates). 7 days post-transduction, RNA was extracted using the Zymo DirectZol RNA purification kit with DNase I treatment following the manufacturer's protocol (Zymo R2052). We divided RNA into 1ug aliquots, using a separate aliquot for each targeted locus. For each locus, the 1 μg of total RNA was reverse transcribed with SuperScript IV Reverse Transcriptase (Thermo Fisher 18090200) to generate cDNA following manufacturer instructions. The anchored poly dT primer for reverse transcription contained a Nextera Read 2 handle (NexteraR2-T18VN; Table S3). After reverse transcription, RNA hydrolysis was performed using RNase H (Thermo Fisher 18021071) and then samples were purified using a DNA Clean & Concentrator-5 kit (Zymo D4013). In the first locus-specific PCR, purified cDNA was amplified with KAPA HiFi HotStart ReadyMix (Kapa Biosystems 07958935001) using the locus-specific primer (Table S3) and adds a Truseq Read 1 handle and a reverse primer containing a sample-specific barcode (P7-index-NexteraR2) (Table S3) with the following cycling conditions: 98°C for 3 min; 15 cycles of 98°C for 20 s, 65°C for 20 s and 72°C for 1 min 20 s; followed by 72°C for 5 min. After amplification, samples were 1X SPRI purified. In the second PCR to prepare samples for next-generation sequencing, the purified product from the previous PCR was amplified with KAPA HiFi HotStart ReadyMix (Kapa Biosystems 07958935001) using a forward PCR primer add sequencing adaptors (P5-TruseqR1; Table S3) and a reverse P7 PCR primer (Table S3) with the following cycling conditions: 98°C for 3 min; 15 cycles of 98°C for 20 s, 60°C for 20 s and 72°C for 1 min 20 s; followed by 72°C for 5 min. Samples were sequenced using an Illumina MiSeq (300 cycle kit), allocating 200 cycles for R1.

In order to identify cleavage positions at single-nucleotide resolution, we first assigned reads to a sample and gene by matching the sample-specific barcode, allowing 1 bp mutation. We then searched for polyA tails in these reads (18 A nucleotides with up to 3 mutations). If a polyA tail was found, this was evidence of proximal cleavage and if no polyA tail was encountered, this was evidence of distal cleavage. For both proximal and distal reads, we additionally required that the last 20bp prior to the cleavage site or end of read were consistent with the expected genomic sequence. This approach enables us to quantify proximal polyA site usage at each locus, but due to limitations of Illumina sequencing, can potentially be biased against capturing transcripts that use the distal site (as they will have larger insert size). We therefore proceeded only with seven out of ten loci where we were able to detect at least 10% of reads showing evidence of distal cleavage.[112] For these loci, we assigned the cleavage site as 1 bp upstream of the most abundant position where we observed a polyA tail in our NT control samples. We compared this position to our inferred cleavage site in CPA-Perturb-seq (Figure S1D).

To quantify the fraction of proximal site polyA usage at each locus, we divided the number of reads where we observed a polyA tail to the total number of filtered reads, as described above. We then calculated the difference in proximal site usage between NUDT21 and NT samples for each locus. We also calculated a similar metric from our CPA-Perturb-seq data by summing counts across all cells receiving NT and NUDT21 guides at the same loci. We observed high consistency between our estimates of NUDT21 perturbation strength in both 3′ RACE and CPA-Perturb-seq data (Figure S1G).

Comparison to A-seq and PAPERCLIP

To compare our polyA site quantifications with previously developed gold-standard methods for polyA site usage (Figures S1H and S1I), we leveraged studies that have profiled polyA site usage in HEK293FT cells with A-seq and PAPERCLIP technologies.[51,52] Fastq files were downloaded using sratoolkit (SRA ID's SRR453410/SRR453411 for A-seq and SRR1810991/SRR1810991 for PAPERCLIP). We used Cutadapt 4.0 to retain reads with sequencing quality of at least q30 and trim both adapters and long polyA

stretches.[113] After collapsing duplicate reads using fastx_collapser, we aligned to GRCh38 with STAR.[114] We quantified polyA site usage for the same regions as in our single-cell data, using featureCounts on the same GFF file as generated above.[115]

First, we examined reproducibility between replicates for NT cells in our single-cell data and control samples for A-seq and PAPERCLIP. To focus on polyA sites that fell within genes that were adequately captured and quantified, we removed genes that fell into the bottom 10% of expression for each experiment. For all replicate comparisons, we considered polyA sites that fell within genes that were adequately captured in both replicates and calculated the percent polyA site usage for each polyA site within a gene for each replicate. We then calculated the Pearson correlation of the percent polyA site usage between replicates. For visualization, we calculated a 2D bivariate normal kernel density estimation, as implemented in the MASS package. We used the same procedure to assess the correlation across technologies, considering polyA sites that fall within genes adequately captured with both technologies.

Mixscape analysis and visualization

We have previously developed a computational toolkit, Mixscape,[48] to analyze and interpret Perturb-seq datasets. Mixscape aims to address confounding sources of heterogeneity in these datasets, and to identify cells that may have received a sgRNA but do not exhibit strong molecular evidence of successful perturbation. We use the polyA counts matrix as input to Mixscape analysis. Briefly, for each regulator, Mixscape considers all cells receiving a sgRNA targeting that regulator, and models them as a Gaussian mixture model with two components. The parameters of the first component are constrained to reflect the distribution of NT control cells. The second learned component should therefore reflect successfully perturbed cells. For each cell that receives a sgRNA, Mixscape calculates a posterior probability for each component. Cells that were classified (posterior probability >0.5) as belonging to the first component are similar to NT control cells (“non-perturbed”), and therefore discarded from further analysis.

In order to visualize our dataset, we used the MixscapeLDA function. All perturbed and NT cells were used as input for this analysis. Linear discriminant analysis (LDA) is a supervised dimensional reduction technique that attempts to find a low-dimensional subspace that maximally discriminates between different groups (‘perturbations’) in our data. As previously described,[48] to prevent overfitting during this procedure, we reduce the dimensionality of our dataset to 360 features prior to running LDA (10 projected principal components × 36 perturbations). Thirty-six components returned from LDA are used as input for 2D visualization in UMAP (Figure 1D).

Single-cell metric for polyA site usage

We sought to develop a single-cell metric for polyA site usage that adjusts for total abundance of a gene in order to prioritize the identification of relative changes in polyA site usage. We focus our analyses on genes that contain more than one polyA site, and leverage NT cells in order to estimate a background distribution describing the usage (counts) of each polyA site. We utilize the Dirichlet-multinomial distribution, allowing us to model overdispersion (relative to the multinomial distribution) that frequently arises in the context of scRNA-seq, due to both biological and technical heterogeneity. The parameters of the Dirichlet multinomial include both the total number of reads observed across all polyA sites, as well as a vector α that yields a probability distribution parameterizing the relative usage of different polyA sites.

Our approach consists of three steps, as described in detail below. (1) Fitting Dirichlet-multinomial parameters for background polyA site usage within each gene individually, using NT cells (2). Regularizing variance estimates across similar polyA sites (3). Comparing polyA sites observed in each perturbed single cell to this model and calculating model residuals (polyA residuals).

Fitting Dirichlet-multinomial parameters

We define the following:

X_{ijk}: number of counts at polyA site k ($1 \leq k \leq K$) for gene j in cell i

n_{ij}: total number of reads in cell i across all polyA sites in gene j $n_{ij} = \sum_{k=1}^{K} X_{ijk}$:

$\alpha_j \in \mathbb{R}^K$: vector of Dirichlet parameters for each of K polyA sites, where each $\alpha_{jk} > 0$

Then, for each cell i, we model the counts for all polyA sites within gene j using the Dirichlet-multinomial distribution.

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

For each gene, we independently estimate the Dirichlet parameters, $\widehat{\alpha}_j$, using count data from 4,336 non-targeting control cells. We obtained parameter estimates using the MGLMfit function in the MGLM package[116] (dist = “DM”), which obtains maximum likelihood parameter estimates using iteratively reweighted Poisson regression.[117]

Once we have obtained the parameter estimates, $\widehat{\alpha}_j$, we can calculate the expected counts falling into each polyA site, as well as the variance of these counts under our background model:

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

$$Var(X_{ijk}) = n_{ij}\frac{\alpha_{jk}}{\sum_{k'=1}^{K}\alpha_{jk'}}\left(1 - \sum_{k'=1}^{K}\alpha_{jk'}\right)\left(\frac{n_{ij}+\sum_{k'=1}^{K}\alpha_{jk'}}{1+\sum_{k'=1}^{K}\alpha_{jk'}}\right) \quad \text{(Equation 2)}$$

We calculate these values for each polyA site in each NT cell.

Regularizing variance estimates

As has been repeatedly observed for both bulk and scRNA-seq data, individual variance estimates obtained from generalized linear models with overdispersed error distributions can be noisy and subject to overfitting.[56,58,59] Therefore, after obtaining estimated variances for our background model, we regularize these estimates across similar polyA sites in order to increase robustness.

We reasoned that polyA sites could be considered similar not only if they were utilized at a similar relative abundance within a gene, but also if the overall gene abundance was similar. We therefore regularize across two dimensions, the expected value of the number of reads falling into each polyA site, $E(X_{ijk})$, as well as the total number of reads that fall into polyA sites in that gene, n_{ij}.

To regularize, we first define a 30 × 30 non-overlapping grid with evenly spaced intervals from the minimum of each dimension to the 99th percentile in NT cells. Each bin is defined by a range of evenly spaced parameter values (intervals) for each of the two dimensions. After defining this grid, we perform the following procedure for each NT cell: we assign each of the polyA sites to one of the bins based on two dimensions, $E(X_{ijk})$ and n_{ij}. For each polyA site, we add estimated variance from Equation 2 to the bin. After completing this procedure for all NT cells, we sample 10,000 observations (polyA sites) from each bin if the number of observations in that bin exceeds 10,000. Then, we use a multivariate kernel regression estimator, as implemented in the kreg function in the gplm package (https://cran.r-project.org/web/packages/gplm/index.html). This calculates a smoothed variance, at the midpoint of each point of the 30 × 30 grid. We use this midpoint as the regularized variance ($\widehat{Var}(X_{ijk})$) for each polyA site falling within that bin.

Calculating polyA residuals for all cells

Given our background model estimated from NT control cells, we can now ask the following questions for any single cell *i* in our dataset: given the observed read depth n_{ij} observed for this cell at gene *j*:

(1) Under the background model, what is the expected value of the counts at each polyA site
(2) Under the background model, what is the variance for these counts
(3) How do the observed values for this cell compare with the expected values?

The expected value for each site can be obtained from Equation 1 above. The variance under the background can be obtained by identifying the correct bin for each polyA site (based on the two dimensions $E(X_{ijk})$ and n_{ij}, and obtaining the regularized variance estimate for that bin. If the dimensions for the polyA site fall outside of the grid defined in the previous step (defined by the 99th percentile in NT cells), we use Equation 2 to calculate the variance. We also set a minimum regularized variance of 0.1 for all polyA sites.

To compare the observed counts from each cell to the expected values under the background model, we calculate a Pearson residual (polyA residual), $\hat{r}_{ijk}$, representing a standardized distance between the observed and expected counts.

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

We calculated two sets of Pearson residuals: one based on tandem sites located in the 3′ most exon, and one on all intronic sites and tandem sites in the 3′ most exon. This allowed for the identification of both tandem and intronic alternative polyadenylation events, as detailed in model-based tests for differential polyA site usage.

By calculating residuals at the polyA site level (instead at the gene level), our procedure can flexibly handle diverse cases, for example, when there are >2 polyA sites in a single 3′ UTR, or when a gene exhibits differential polyadenylation at both intronic and tandem polyA sites. However, we note that the polyA residuals at multiple sites in the same gene are not independent (i.e., if one polyA site exhibits high relative usage and receives a positive residual within a single cell, other polyA sites will likely decrease in relative use and receive negative residuals). For this reason, when we perform downstream analyses such as clustering or principal components analysis (which often assume independence across features) using polyA residuals, we only consider a single polyA site per gene.

Defining modules based on CSTF/CPSF responsiveness

We aimed to first identify a set of sites that were responsive to NUDT21 perturbation (shortening) but could in principle result in lengthening in response to other perturbations (Figure 4). We identified 1,583 polyA sites located in the last exon whose usage was decreased in response to NUDT21 perturbation and increased in response to RBBP6 perturbation, using a threshold of $q < 0.05$ and a percent change compared to NT that was greater than 1%. These sites represent predominantly distal sites whose usage decreased in CF-I perturbation and increased in RBBP6 perturbation.

Then, we identified a set of sites that were upregulated in both CSTF and CPSF perturbations (using the complex-based linear model described below), representing 323 genes. These genes compose module A, where CPSF/CSTF perturbation acts in the opposite direction of NUDT21 perturbation. We also identified 149 polyA sites composing module B, whose usage decreased in CSTF and CPSF perturbation, indicating that NUDT21 perturbation and CPSF/CSTF perturbation induced 3′ UTR shortening at these genes.

APARENT-Perturb model and interpretation

Data processing

The data was aggregated to pseudo-bulk level by grouping cells by perturbation gene. The polyA sites of each gene were intersected against the coordinates of annotated sites in PolyADB V3.[7] Only sites within the 3′ UTR were retained. Sites without a well-defined core hexamer motif (at most two substitutions away from the canonical AAUAAA motif) were removed. Finally, genes with less than two retained sites, or more than 10 sites, were thrown out. The resulting filtered data consisted of 5,267 genes. For each gene and perturbation, we used the isoform read counts to estimate the relative usage (proportion) of each polyA site.

Architecture and training

The neural network follows a residual architecture with 6 layers of dilated convolutions (32 filters per layer, each filter is 3 positions wide, the dilation rate is doubled per layer) and one initial convolutional layer with filter size 5 and no dilation before the residual layers, taking as input 205 bp of sequence in a one-hot-encoded format centered on the 3′ cleavage site and outputs residual logit scores for the NT condition and the 10 chosen perturbations. The architecture differs slightly from the tissue-specific multi-PAS model proposed in APARENT2.[43] Specifically, the neural network does not only output one logit score per sequence and perturbation, but instead it outputs three scores per perturbation: a proximal site score, a middle-site score and a distal site score. This allows the model to learn different *cis*-regulatory rules of polyA site usage depending on site type. The model is replicated and shared across all polyA sites of a given gene (for a maximum of 10 sites), such that the model predicts a tensor of perturbation response logits $S \in \mathbb{R}^{10\times11\times3}$ given an input tensor of polyA site sequences $X \in \{0,1\}^{10\times205\times4}$. The response logits S are used in combination with baseline APARENT2 logits $A \in \mathbb{R}^{10}$ and the log distance between sites, $D \in \mathbb{R}^{10}$, to predict the final proportion of usage for each polyA site and perturbation, $\hat{Y} \in [0,1]^{10\times11}$:

$$\hat{y}_{i,j} = \frac{\mathbb{1}_{\{\text{PAS } i \text{ exists}\}} \times \exp(\text{Score}(i,j))}{\sum_{t=1}^{10} \mathbb{1}_{\{\text{PAS t exists}\}} \times \exp(\text{Score}(t,j))}$$

Where:

$$\text{Score}_{i,j} = \left[\sum_{k \in \{p,m,d\}} \mathbb{1}_{\{\text{PAS } i \text{ is type } k\}} \times w_k^{(\text{score})} \times (a_i + s_{i,j,k}) \right] + w^{(\text{distance})} \times d_i + w_{i,j}^{(\text{bias})}$$

And:

i: Site # in 3′ UTR
j: Perturbation #
k: Proximal, middle, or distal output
a_i: APARENT2 score for site i
d_i: Cumulative log-distance (bp) from proximal-most site
$s_{i,j,k}$: Predicted score of perturbation j at site i for type k (trainable)
w: Regression weights (trainable)

Similar to the training procedure described in APARENT2, we first fit the regression weights w while keeping the perturbation-specific scores $s_{i,j,k}$ frozen at zero. During this phase, we minimize the KL-divergence between measured proportions Y and predicted proportions $\hat{Y}$. After convergence, we unfreeze the scores $s_{i,j,k}$ and fit the weights of the underlying neural network using a combination of a KL-divergence loss and a margin error between measured differences ($Y_{\text{Perturb}} - Y_{\text{NT}}$) and predicted differences ($\hat{Y}_{\text{Perturb}} - Y_{\text{NT}}$). The model was trained with Keras using the Adam optimizer[118] and training stopped once the loss started to increase on held-out validation data.

In silico saturation mutagenesis and motif finding

Attribution scores were generated for the polyA sites of all 5,267 genes and for each perturbation using a windowed In-silico Saturation Mutagenesis (ISM) approach.[80,81] By sliding a 5 nt wide window over each sequence, we randomly and uniformly mutate all 5 bases within the window and record the mean difference in predicted logit scores with respect to the wild-type sequence for 8 independent samples. The resulting mean log-odds ratio is taken as the importance score for the base at the center of the window. Whenever the importance scores for a particular perturbation are analyzed, we always subtract the importance scores of the NT condition from the scores of the perturbation.

To cluster the importance scores and discover motifs, we ran TF-MoDISco[82] on the scores of each perturbation with the following settings: sliding window = 8, flank size = 5, max seqlets = 40,000, FDR = 0.1 and # mismatches = 1). TF-MoDISco was executed separately on the negative- and positive part of the importance scores.

Motif insertion analysis and pairwise ablations

Dual motif insertion simulations were performed on a set of 64 wild-type polyA site sequences as backgrounds. These sites were chosen from all polyA sites that do not already contain instances of the motif of interest. The sites were uniformly sampled from this set unless otherwise specified (e.g., in some analyses we biased the selection to GC- or AT-rich contexts). For each wild-type sequence, we predicted the residual score $r^{(WT)} = s_{\text{Perturb}} - s_{\text{NT}}$ for the perturbation of interest. We then inserted motif A, motif B or motif A and B at every possible combination of positions (i and j) and generated the corresponding predictions $r^{(A)}_{i,j}$, $r^{(B)}_{i,j}$, $r^{(A \text{ and } B)}_{i,j}$. Given these

quantities, we estimated the odds ratio of epistasis at positions *i* and *j* as: epi = $\exp((r^{(A\text{ and }B)}_{i,j} - r^{(WT)}) - ((r^{(A)}_{i,j} - r^{(WT)}) + (r^{(B)}_{i,j} - r^{(WT)})))$. Finally, by grouping pairs of positions by their distance ($|i - j|$), we estimated the 10th, 50th and 90th percentile of odds ratios for each distance.

Pairwise motif ablations were estimated using an identical formula, except now motifs are ablated (mutated) from wild-type sequences where they already exist. Motifs were ablated by replacing the seqlet identified by MoDISco with uniformly random bases for 32 independent samples, unless the motif occurred over highly conserved positions (e.g., the core hexamer), in which case the motif was randomly exchanged for specific hand-crafted subsequences (e.g., weaker version of the hexamer motif).

Massively parallel reporter assay experiment

MPRA library cloning

We based our MPRA library design on the APA Reference Library (Addgene #193784), which was used to validate the APARENT2 model.[43] The construct enables the profiling of thousands of polyA site sequences, which are inserted into the proximal site, but also enables the user to choose (via insertion) the sequence context at the distal site. We chose to test our MPRA library using five different distal site contexts, with varying strengths, and included a distal site barcode which enabled us to know which context was relevant for each MPRA sequence read.

To do this, we first separately cloned five distal sequences (bGH, CCT6A, TMEM237, TMEM106C, or CDK1; Table S5) into the distal site of the construct. We ordered the five sites as gBlocks from IDT with the following structure: PCR-handle(containing Bsp1407I cut site)Nextera-Read2distal-site-barcodeEsp3I-cut-sitefiller-sequencevariable-proximal-siteSalI-cut-sitePCR-handle (Table S5). gBlocks were amplified with forward (5′-GGCATGGACGAGCTGTAC-3′) and reverse (5′-CCGAAAAGTGCCACCTGAC-3′) primers and Q5 polymerase (NEB M0492S) using the following PCR program: 98°C for 30 s; 25 cycles of 98°C for 10 s, 60°C for 20 s and 72°C for 20 s; followed by 72°C for 3 min. The amplified gBlocks were ligated into the Bsp14071 and SalI-digested APA Reference Library vector using a Gibson Assembly reaction (NEB E2611S), following manufacturer recommendations.

We designed 3,802 proximal sequences for testing, but also included barcode replicates for a subset of sequences, leading to a total of 6,113 sequences (see MPRA Construct Design below; Table S5). We ordered an oligonucleotide pool from Twist Biosciences containing these sequences, flanked by Esp3I restriction enzyme cut sites. We cloned this pool into each of the five distal site constructs. 20 ng of the oligonucleotide pool was amplified with forward (5′-TAGAAGGTCTATGTTCGCCA-3′) and reverse (5′-TAAC GAGTCCTAAACGGGAT-3′) primers and KAPA HiFi HotStart ReadyMix (Kapa Biosystems 07958935001) using the following PCR program: 98°C for 30 s; 9 cycles of 98°C for 10 s, 60°C for 15 s and 72°C for 15 s; followed by 72°C for 3 min. The amplified oligonucleotide pool was 1.4X SPRI purified, followed by Esp3I (Thermo Fisher FD0454) digestion and an additional 2X SPRI purification. Each of the five distal site plasmids were all digested with Esp3I (Thermo Fisher FD0454) and then dephosphorylated by incubation with FASTAP (Thermo Fisher EF0654). The digested proximal site reporter insert was split into five equal amounts and ligated into each of the digested distal site plasmids using T7 DNA ligase (NEB M0318S). After cloning, the five distal site plasmid pools were combined at equal amounts to generate a final MPRA library.

Library transfection into

CRISPRi-expressing HEK293FT cells were transduced with lentivirus for CSTF3, NUDT21, NT (two guides and 2 replicates each for a total of twelve samples) and 8 μg/mL polybrene (EMD Millipore TR-1003-G). After three days of selection for guide positive cells with 1 μg/mL puromycin (Thermo Fisher A1113803), 1,000,000 cells were seeded per well in a 6-well plate 16–24 h prior to transfection. The MPRA library was transfected into cells using Lipofectamine 3000 (Thermo Fisher L3000008) using 2.5 μg plasmid library per well.

RNA extraction

48 h post-transfection (7 days after original transduction with sgRNAs), RNA was extracted using the Zymo DirectZol RNA purification kit with DNase I treatment following the manufacturer's protocol (Zymo R2052). mRNA was isolated from 5 μg total RNA per sample using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB E7490S).

MPRA library preparation and sequencing

The resulting mRNAs were reverse transcribed with SuperScript IV Reverse Transcriptase (Thermo Fisher 18090200) to generate cDNA following manufacturer instructions. The anchored poly dT primer for reverse transcription contained a Truseq Read 1 handle, an 8 bp sample barcode, and a unique molecule identifier (UMI) (TruseqR1-barcode-UMI-T18VN, see Table S5 for primer sequences). After reverse transcription, RNA hydrolysis was performed using RNase H (Thermo Fisher 18021071) and then samples were purified using a DNA Clean & Concentrator-5 kit (Zymo D4013). Purified cDNA was amplified with KAPA HiFi HotStart ReadyMix (Kapa Biosystems 07958935001) using a forward primer containing additional Illumina adaptor sequences (P5-TruseqR1; Table S5) and reverse primer containing a sample-specific barcode (P7-index-NexteraR2; Table S5). Amplification was monitored with qPCR so that the reaction could be stopped early to minimize PCR bias with the following cycling conditions: 98°C for 3 min; 15 cycles of 98°C for 20 s, 60°C for 20 s and 72°C for 1 min; followed by 72°C for 5 min. After amplification, samples were 1X SPRI purified. The MPRA library was sequenced using an Illumina NextSeq Mid-Output (300 cycle) kit: allocating 16 cycles to Read 1, and 200 cycles to Read 2.

Massively parallel reporter assay: Construct design and computational analysis

We designed sequences to validate APARENT-Perturb predictions as follows.

Modifying CSTF perturbation response

We selected 110 CSTF3-responsive sequences, and 110 nonresponsive sequences. For both sets, we divided sequences between those natively located at proximal or distal polyA sites. The perturbation-responsive sequences were selected in two steps: First, we picked the top 4% most responsive sequences based on the APARENT-Perturb CSTF3 model predictions, after which we sorted the resulting subset in descending order of measured perturbation response and chose the top sequences. The nonresponsive set was chosen as sequences exhibiting minimal predicted responsiveness in the APARENT-Perturb CSTF3 model.

We next created sequence alterations in order to modulate responsiveness to CSTF3 perturbation. To reduce responsivity, we used the APARENT-Perturb CSTF3 model to identify the 9 bp window with the strongest mode of ISM score at each locus. We then replaced this window with random sequence (probability = A = C = G = T = 25%), which we refer to as a "shuffle". Each shuffle was repeated 5 times with new random sequences. To design sequences with increased CSTF3 responsivity, we initialized each starting sequence with one of the CSTF3-responsive wild-type sites and used MCMC sampling to "improve" them with at most 10 or 20 mutations. We restricted the design procedure from altering the core hexamer. We further restricted the procedure from creating overly A-rich repeat regions (no 7-mer with 6 or more A's), which are problematic to quantify experimentally due to internal priming.

We also classified WT sequences based on whether they contained a canonical or non-canonical CSTF binding site. We defined polyA sites as canonical if the 9bp region with strongest mode of ISM score for the CSTF3 model contained AT-rich (NTTTT/TNTTT/TTNTT/TTTNT/TTTTN) or GT-rich (GTTGT/TGTGT/GTGTT) downstream sequence element.[89] Wild-type sequences where the region chosen for shuffling did not contain one of these elements were considered non-canonical.

Altering module/B gene proximal site bias

To test our hypothesis that differential strengths of CSTF proximal site binding drives differences in the proximal bias of module A/B genes, we started with the proximal sites of 49 module A genes and 47 module B genes. For module A genes, we shuffled the GT-rich downstream region with the highest ISM score (as before, we replaced a 9bp window with random nucleotide content). For module B genes, we inserted a GT-rich motif (chosen randomly from TGTGTGTGT, TGTGTGTCT, TGTGTCTGT, TGTCTGTGT, TCTGTGTGT, TGTGTTTTT, TTGTGTTTT, TTTGTGTTT, TTTTGTGTT, TTTTTGTGT, TGTCTTTTT, TTGTCTTTT, TTTGTCTTT, TTTTGTCTT, or TTTTTGTCT). The insertion was placed in the downstream region (after the core hexamer and at most 70bp downstream), and we used the APARENT-Perturb CSTF3 model to identify the position where the motif insertion was predicted to elicit a maximal response.

Testing for epistasis between pairs of motifs

To test for epistasis between NUDT21 motifs, we started with 19 neutral sequences that did not contain any TGTA motifs. We performed single and dual insertions of TGTA with AT-rich flanks (TTTGTAAT) and GC-rich flanks (GGTGTAGC). We performed this insertion at 4 different distances between motifs (0bp, 5 bp, 20 bp, and 40 bp).

To test for epistasis between the core hexamer and downstream GT-rich elements, we started with 50 sequences that contained a weak core hexamer sequence (one of ATTAAA, AGTAAA, ACTAAA, TATAAA, GATAAA, or CATAAA) and no downstream GT-rich motif (no occurrence of GTG, TGT, GTCT, TGTC, TGTTGT, GTTT, TTTT, TTAT, GTTT, or TTTG). We replaced the weak hexamer with a strong hexamer (AATAATAAATAA). In addition, we inserted a GT-rich motif (TTGTGTGTT) downstream of the core hexamer at various distance offsets (0, 5, 10, 20, 40). We performed these modifications individually and simultaneously to test for potential epistasis.

Quantifying polyA site usage from MPRA sequencing

We demultiplexed the MPRA data using the barcode sequences that were synthesized into the oligonucleotide pools and cloned into the MPRA construct. The barcodes in Read 1 identified the experimental condition (i.e., replicate and sgRNA identity), as well as a unique molecular identifier. The barcodes in Read 2 identify the proximal and distal site contexts that gave rise to the mRNA molecule being sequenced. The distal site barcodes were assigned using the first 4–8 bp of read 2 (see Table S5). The next 20 bp after the distal site barcode were used to assign reads to a proximal site (one of our 6,113 polyA site constructs, each with a unique 20 bp barcode). Only reads that matched distal and proximal barcodes within 1 bp each were retained. We collapsed all reads with the same UMI, proximal, and distal site barcodes (allowing 1 bp mutations in the UMI sequence).

In our MPRA sequencing data, if an mRNA molecule is cleaved at the proximal site, we will observe a stretch of A nucleotides at the cleavage site in the sequence read. If the mRNA molecule is distally cleaved, the full read will consist of the 3′ UTR sequence (we will not observe a polyA stretch). Therefore, for each read, we classified it as "proximal" if it contained a polyA stretch (18 A's with up to 2 bp mismatches). Reads that did not contain a polyA stretch were classified as distal. We only retained reads where the last 20bp prior to the cleavage site or end of read were consistent (within a Levenshtein distance of 3) with the expected genomic sequence. We then obtained a counts matrix of proximal vs. distal reads for each sample and distal site. We pooled counts from samples profiled with independent guides across the same target gene, as we observed high reproducibility across independent guides (Figure S6C). We also summed counts across barcode replicates and independent shuffles of the same locus.

In total, our MPRA profiled mRNA molecules originating from 30,565 (6,113 proximal site × 5 distal sites) contexts. We obtained a median of 267 UMI for each mRNA molecule in an experimental condition, enabling us to accurately quantify proximal vs. distal polyA site usage. As quality control, we removed rare sequences where we obtained less than 100 total UMI (summing across repeat shuffles and guide replicates), or where we observed exclusively proximal or distal reads. For sequences that passed QC, we calculated the log odds of proximal site usage: $\log_2$(Proximal reads/Distal reads). We averaged this value across the two biological replicates. To

compare the measured log odds with APARENT-Perturb predictions in both NT and CSTF3 perturbed cells (Figure 6B), we used APARENT2 estimates for NT samples and summed these estimates with predicted log-odds ratio for CSTF3 from APARENT-Perturb to obtain estimates of proximal site usage in CSTF3 perturbed cells.

In order to visualize how modifications of native polyA sites effects their usage (Figures 6E–6H and 6K), we took all assigned/filtered reads (as described above), and plotted how frequently we observed a polyA stretch at each position in the construct (0–164 bp). Spikes in the visualization track indicate the location of proximal cleavage sites, and the height of the track corresponds to the percentage of mRNA molecules that are proximally cleaved. Each plot shows all filtered reads in the context of one distal site (bGH for Figure 6E, CCT6A for Figure 6H, and CDK1 for Figure 6K).

Analysis of genome-scale Perturb-seq dataset

We processed the scRNA-seq gene expression raw data from the Genome-scale Perturb-seq dataset[47] (K562 KD8) using CellRanger (v6.0.0, ''cellranger count''; hg38, ensembl v97). Guide RNA and target gene assignments were extracted from the metadata of the original analysis (https://gwps.wi.mit.edu). We quantified polyA site counts as described previously (quantifying polyA site usage), using the same polyA site GFF file as in the HEK293FT dataset. From this dataset, we extracted cells with sgRNAs targeting 1,280 previously annotated RNA binding proteins,[90] along with NT controls. We kept all polyAsites located within 50bp of a previously annotated site in either the polyAdbv3[7] database. After intersection, we retained sites whose annotated location was either within an intron or the last exon.

We performed differential polyadenylation for each perturbation as previously described for our CPA-Perturb-seq dataset (model-based tests for differential polyA site usage). After linear modeling, we defined differential APA events based on an FDR-threshold of 0.05. To compare these results with K562 cells profiled with CPA-Perturb-seq (Figure 7A), we calculated the fraction of tandem events that were associated with 3′ UTR shortening in each dataset (for overlapping tandem regulators with at least 25 cells in the GPWS dataset). For all perturbations with more than 100 APA events, we calculate a correlation matrix using the same procedure used in Figures 4A and 4B. In order to focus on perturbations that cluster together, we retain regulators that exhibit a correlation of >0.25 with at least one other perturbation and visualize this correlation matrix in Figure 7C. We defined regulator ''modules'' based on correlated perturbation responses in Figure 7C.

To investigate if intronic signatures defined from the CPA-Perturb-seq data replicated for any of GWPS perturbation modules, we repeated differential polyadenylation analysis using the modules we defined in Figure 7C. Then, we used a hypergeometric test to test if the list of intronic polyA sites identified from GWPS significantly overlapped with the list of signature sites from CPA-Perturb-seq. In Figure 7G, the size of dots indicates the p value from the hypergeometric test. The color of each dot indicates the average polyA residual across the intronic polyA sites for each signature, across all cells in each GWPS module. In Figure 7H, we assess the strength of 5′ splice site (donor) sequences for intronic polyA sites. To do this, we used the calculateMaxEntScanScore in the VarCon R package.[119] This package calculates splice site scores using 9 bp of the sequence near the 5′ donor site, which we extracted using the GenomicRanges Bioconductor package.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model-based tests for differential polyA site usage

In order to identify polyA sites whose relative usage changes after each perturbation, we performed differential analysis on the polyA residuals using a linear model. For each site k and each perturbation, we ran a linear model comparing the perturbed cells to NT cells with main effects for each of l guides that target the same gene.

$\hat{r}_{ijk} = \beta_0 + \beta_1 Z_{i1} + \ldots + \beta_1 Z_{il}$,

Z_{il} = 1 if cell i received guide l, 0 otherwise

$Z_{i1} = \ldots = Z_{il} = 0$, when cell i received NT guide

For each site, we only include cells with non-zero values for the residuals in the linear model, representing cells in which at least one read aligned to polyA sites in the same gene was captured. We then prepare a contrast in order to compare all of the guides targeting the same gene to NT cells, as well as comparisons between guides. For instance, if there are 3 guides that target a particular gene, the contrast matrix will be as follows:

$$\begin{pmatrix} -1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{3} & -1 & 1 \\ \frac{1}{3} & 0 & -1 \end{pmatrix}$$

The first column compares NT cells to the average across each guide. The second column compares the first and second guides, and the third column compares the second and third guides. By weighting each guide equally we will downweight observations

where the perturbation strength is not observable across multiple independent sgRNA. In order to test the specified contrasts matrix, we use it as input to the contrasts argument to the lm function in R. The corresponding estimate ("perturbation effect estimate") and *p* value for the contrast comparing NT cells to all guides is then used to assess if a particular site responds to a perturbation.

Linear modeling was performed separately on both the tandem sites and intronic polyA residuals. After running this linear model on all polyA sites, the q-value package (https://github.com/StoreyLab/qvalue) was used to estimate q-values with a false-discovery rate (FDR) of 5%. Genes having at least one tandem site with a q-value less than 0.05 and greater than 5% change in average usage compared to NT cells were classified as tandem alternative polyadenylation events. Similarly, genes with at least one intronic site with a q-value less than 0.05 and at least 5% change compared to NT cells were classified as intronic differential polyadenylation events. For Figure 3A (left), a gene could have both intronic and tandem differential polyadenylation events.

We also utilize the results of the linear model to perform clustering on the alternative polyadenylation perturbation responses across regulators. For regulators that exhibit primarily tandem APA, we first select all genes that show evidence of tandem alternative polyadenylation in any perturbation, and extract the top tandem polyA site (based on the magnitude of the polyA residual) for each gene. For each regulator, we define a vector of perturbation effect estimates (obtained from our linear model contrasts) for each of these polyA sites. These vectors were used as input to hierarchical clustering in Figures 4A and 4B. We also perform a similar procedure for regulators that exhibit primarily intronic APA (Figure S4E), using perturbation effect estimators from the top intronic polyA site for each gene as input to correlation and hierarchical clustering.

In order to intersect changes in polyA site usage with potential changes in gene abundance, we also estimated total RNA levels for each gene with multiple polyA sites. We summed counts across all annotated polyA sites for that gene for each cell, regardless of position within the transcript, and performed standard log-normalization in Seurat. We then performed differential expression to compare gene expression of each perturbation to NT cells, using the Wilcoxon rank-sum test (Figures 2F–2H). Significance was determined using a false discovery threshold of 0.05, as above. Genes were defined as differentially expressed for a perturbation if the log2FC exceeded 0.25 and the q-value was less than 0.05.

To identify a robust set of peaks that respond similarly across members of the same complex (Figure 4C), we ran an additional model testing for the difference between NT cells and guides targeting the same complex (CPSF1-4 and FIP1L1 for CPSF and CSTF1/3 for CSTF).

Linking intronic polyA sites with RNA life cycle

We focused on 8 regulators whose perturbation primarily resulted in increased usage of intronic polyA sites: CDC73, PAF1, CTR9, SCAF8, SCAF4, PABPN1, CPSF3L, and RPAP2. For each perturbation, we identified significant intronic sites that had an increase in average usage >5% compared to NT cells as well as q-value <0.05, as described above. We performed hierarchical clustering of intronic regulators using the same procedure as we define above for tandem regulators, leading us to define two correlated groups of regulators: PAF complex members (PAF1/CTR9/CDC73) as well as CPSF3L/RPAP2. We defined a signature of responsive sites for these two groups (using the intersection of significant sites from each group member), and for the remaining individual regulators (PABN1, SCAF4, SCAF8). For Figure 3D, we plot the top 50 unique sites for each regulator or set of regulators.To understand factors that drive intronic polyA site usage for each regulator, we first calculated a set of features for each intronic site. We identified the intron where each polyA site fell, using the knownCanonical table obtained from the UCSC genome browser (hg38). We calculated the width of that intron, its GC content, the total width of the transcript (subtracting the contribution of that intron), the distance to the next polyA site in the gene, and whether or not it was the first intron. We log2-transformed all widths and distances and used these features as covariates in a linear model. For each regulator, we aimed to predict the perturbation response for each intronic site (represented by its differential polyA site usage coefficient, as defined above), based on these covariates. For Figure 3E, we show the learned t statistic for the covariates in each regulator model.

We further characterized the relationship between distance to the next polyA site and perturbation response at each intronic polyA site. We binned polyA sites into deciles based on the next annotated polyA site in the gene and then calculated the Pearson correlation between distance (log2-distance) and the percent change of the intronic polyA site in PAF1-perturbed cells compared to NT (Figure S4F). We also repeated this analysis for CPSF3L-perturbed cells. To assess if this relationship for PAF1 was unique to sites located in the first intron, we also performed this analysis separately for polyA sites located within first introns, and all other introns (Figure S4G). We also repeated this analysis binning polyA sites into deciles based on GC content and intron width (Figures S4H and S4I).

In Figure 3I, we used previously published data that calculated intron excision speeds for introns in 4,163 genes, 2,212 of which we profiled in our dataset.[62] We downloaded summary data for these introns from the primary manuscript, and after lifting over coordinates from hg19 to hg38, we identified 519 introns that overlapped with the cleavage site of an intronic polyA site with increased usage in one or more perturbations. We calculated the fraction of these polyA sites with slow, medium, or fast intron excision speeds ("thetaClust" in the original dataset).

Polynomial feature regression

Polynomial feature regression was used to validate the epistatic relationships discovered by the neural network directly in the data (Figures S5H–S5K). To validate the epistasis of a given pair of motifs A and B, we constructed independent count features of the occurrence counts of motif A and motif B respectively, as well as binary indicator features of higher-order combinatorial

co-occurrences (e.g., both motif A and B occur *n* times in site *i*). These features were constructed for each polyA site in each gene. To reduce the number of free parameters, the features for all non-distal sites in the same gene were aggregated. Given this feature matrix as input, we performed standard linear regression in scikit-learn[120] to regress the log of the odds ratio of distal polyA site usage in the perturbation of interest relative to the NT condition. To interpret the learned epistatic rules, we inspect the signs and magnitudes of both the individual count feature weights and the combinatorial binary indicator weights. For example, to assess the homotypic relationship between instances of motif A, we perform regression using a count feature of motif A and a binary indicator which is set to 1 only if two or more instances of motif A are present in the sequence. After regression, if the learned weight for the binary indicator is negative it suggests a sub-additive relationship (the regression subtracts from the additive count feature). Conversely, a positive indicator weight suggests a super-additive relationship (the regression adds to the additive count features).

Validating sequence effect on polyA site usage

In order to compare proximal site usage between two conditions (or sequences) in our MPRA library, we calculated the log-odds ratio: the ratio of the two log-odds values. In Figure 6C, we performed a two-sample Wilcoxon test comparing the log-odds ratios (log-odds CSTF3/log-odds NT) of predicted CSTF3-responsive sequences to nonresponsive sequences. To test whether sequence alterations affect perturbation responsivity (Figures 6F and 6G) we calculated the Δ log-odds ratio (i.e., the difference in log-odds ratio between the WT sequence and the corresponding altered sequence), and then performed a one-sample t-test to test if this value was significantly different from 0. Finally, to test for epistasis (Figure 6N), we first calculated the log-odds ratio comparing proximal site usage of the altered sequence to the WT sequence, for each individual/dual insertion. We calculated the epistasis odds ratio: = 2^(LOR(dual insertion) − [LOR(single insertion 1) + LOR(single insertion 2)]). We then calculated the difference in epistasis odds ratio between NT and NUDT21-perturbed cells and performed a one-sample t-test to test if this value was significantly different from 0. The boxplots in Figures 6 and S6 depict the median (horizontal line) and first and third quartiles (lower and upper hinges, respectively). The upper and lower whiskers extend to 1.5 times the interquartile range from each hinge, and outlier points falling beyond the ends of the whiskers were removed for visual clarity.

CellPress

Supplemental figures

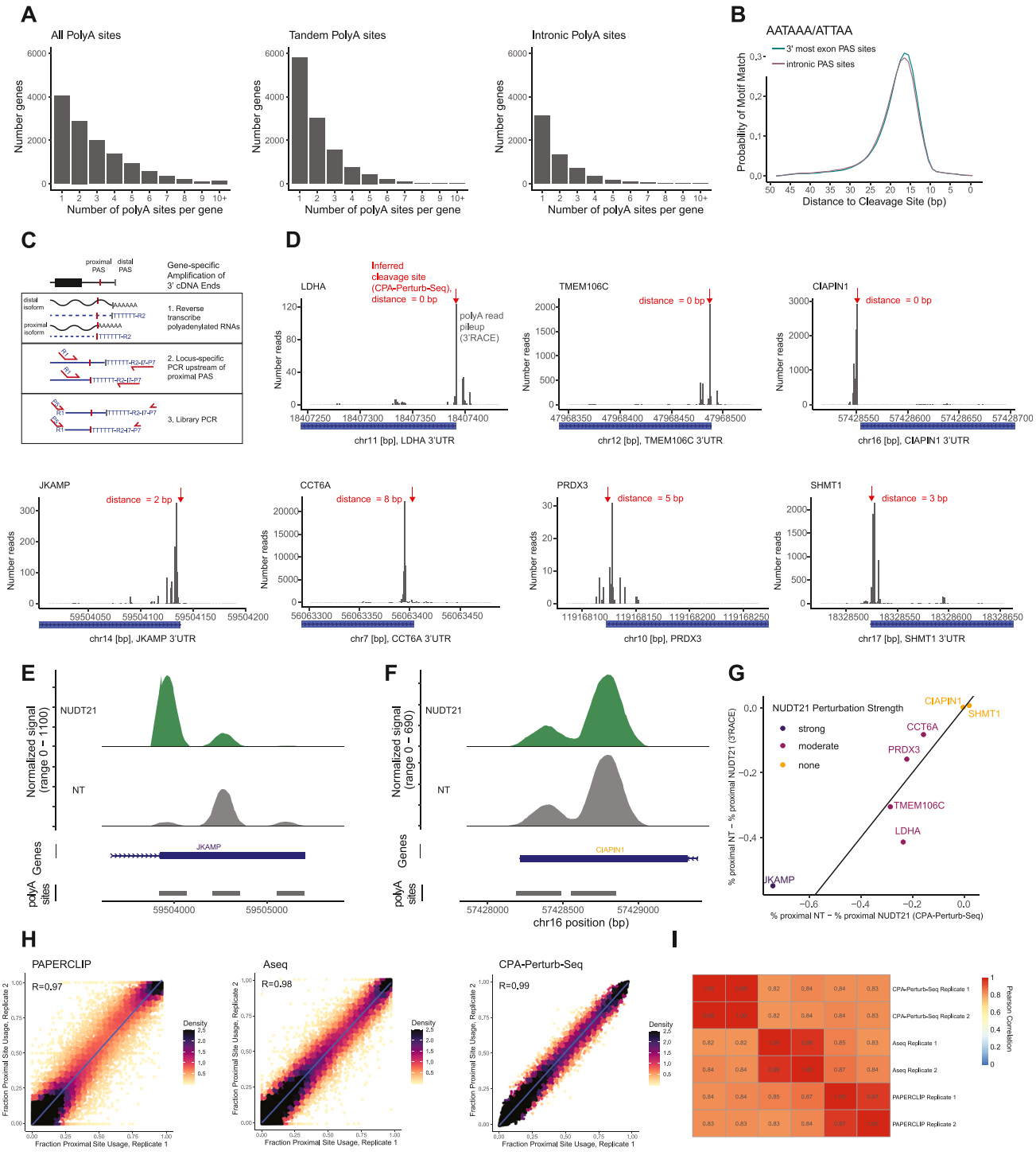


Figure S1. Profiling polyA site usage with 3′ scRNA-seq, related to Figure 1

(A) Histogram showing the number of identified polyA sites (after intersection with polyA_DB v3.2) per gene in the CPA-Perturb-seq dataset.

(B) Positional preference of the canonical cleavage motives AATAAA and ATTAAA relative to the detected polyA sites in terminal exons and introns.

(C) Schematic of gene-specific amplification of 3′ cDNA ends (3′ RACE), used to identify cleavage sites at single-base-pair resolution.

(legend continued on next page)

(D) Visualization of frequency of reads (y axis) containing a polyA sequence (evidence of proximal cleavage) obtained from 3′ RACE with Illumina sequencing. Blue bar shows the predicted transcript, and red arrow indicates the inferred polyA site from CPA-Perturb-seq.
(E and F) Read coverage plot from CPA-Perturb-seq from the JKAMP (left) and CIAPIN1 (right) loci. Each track represents a pseudobulk average of cells, grouped by target gene perturbation. JKAMP exhibits a strong increase in proximal site usage upon NUDT21 perturbation, and CIAPIN1 does not exhibit changes in polyA site usage upon NUDT21 perturbation.
(G) Change in proximal site usage comparing NT to NUDT21 knockdown in CPA-Perturb-seq (x axis) to 3′ RACE (y axis) for 7 genes. Points are colored by NUDT21 perturbation strength, as measured in CPA-Perturb-seq.
(H) Scatterplots showing reproducibility across replicates when profiling polyA site usage in HEK293FT cells at baseline using different technologies. Each point represents percentage isoform usage in replicate 1 (x axis) vs. replicate 2 (y axis). Points are shaded according to kernel density estimation (clipped at 2.5), and the Pearson correlation between replicates is calculated.
(I) Pearson correlation matrix showing quantitative transcriptome-wide agreement across three different technologies used to profile polyA site usage.

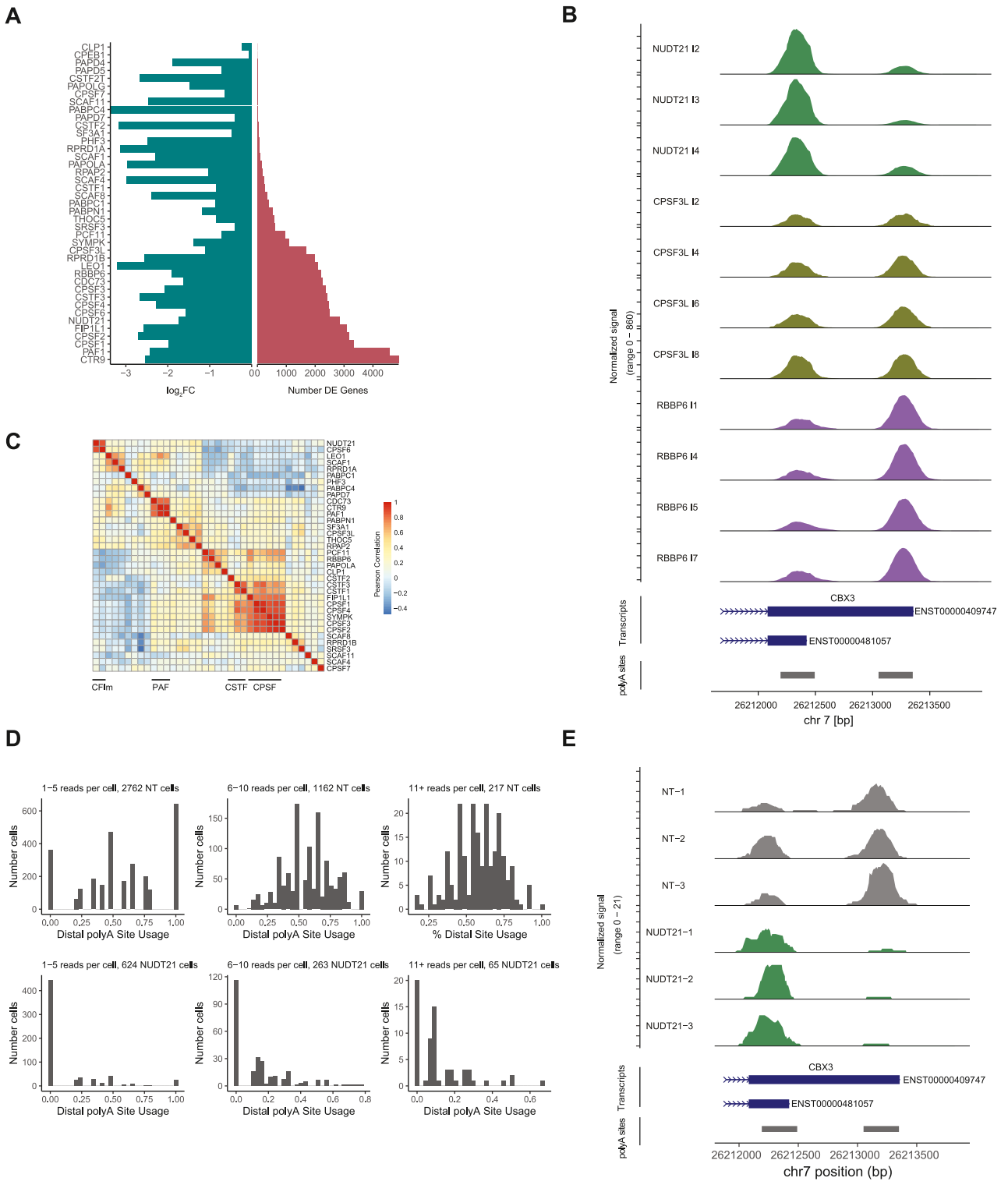


Figure S2. Perturbation responses in the CPA-Perturb-seq dataset, related to Figure 1

(A) Log-fold change of the target gene observed in the Perturb-seq data (left) vs. the number of differentially expressed genes for that perturbation (right). In cases where either no or few differentially expressed (DE) genes were detected, Mixscape will classify all cells as "non-perturbed." This can occur in cases where the target perturbation was unsuccessful (i.e., CLP1) but also in cases where the target gene was downregulated but no global effect was observed (i.e., PABC4).

(legend continued on next page)

(B) Read coverage plot at the CBX3 locus. Each track represents a pseudobulk average of cells, grouped by their perturbation and separated by their individual sgRNA.
(C) Hierarchical clustering of the polyA site/regulator count matrix. Each column represents pseudobulk average of single cells that received the same target gene perturbation, after subtracting the pseudobulk average of NT control cells.
(D) Distribution of fraction of distal polyA site usage within single cells for CBX3 locus, split by total number of reads for CBX3 within each cell, for both NT cells (top) and cells with NUDT21 perturbation (bottom). Cells with low coverage at this locus (1–5 reads, left) show a primarily bimodal distribution between proximal and distal site usage, but cells with sufficient sequencing depth (middle, right) show that polyA site usage is heterogeneous even with individual cells.
(E) Read coverage at the CBX3 locus for 3 NT cells (gray) and 3 NUDT21-perturbed cells (green), illustrating heterogeneity in polyA site usage within individual cells.

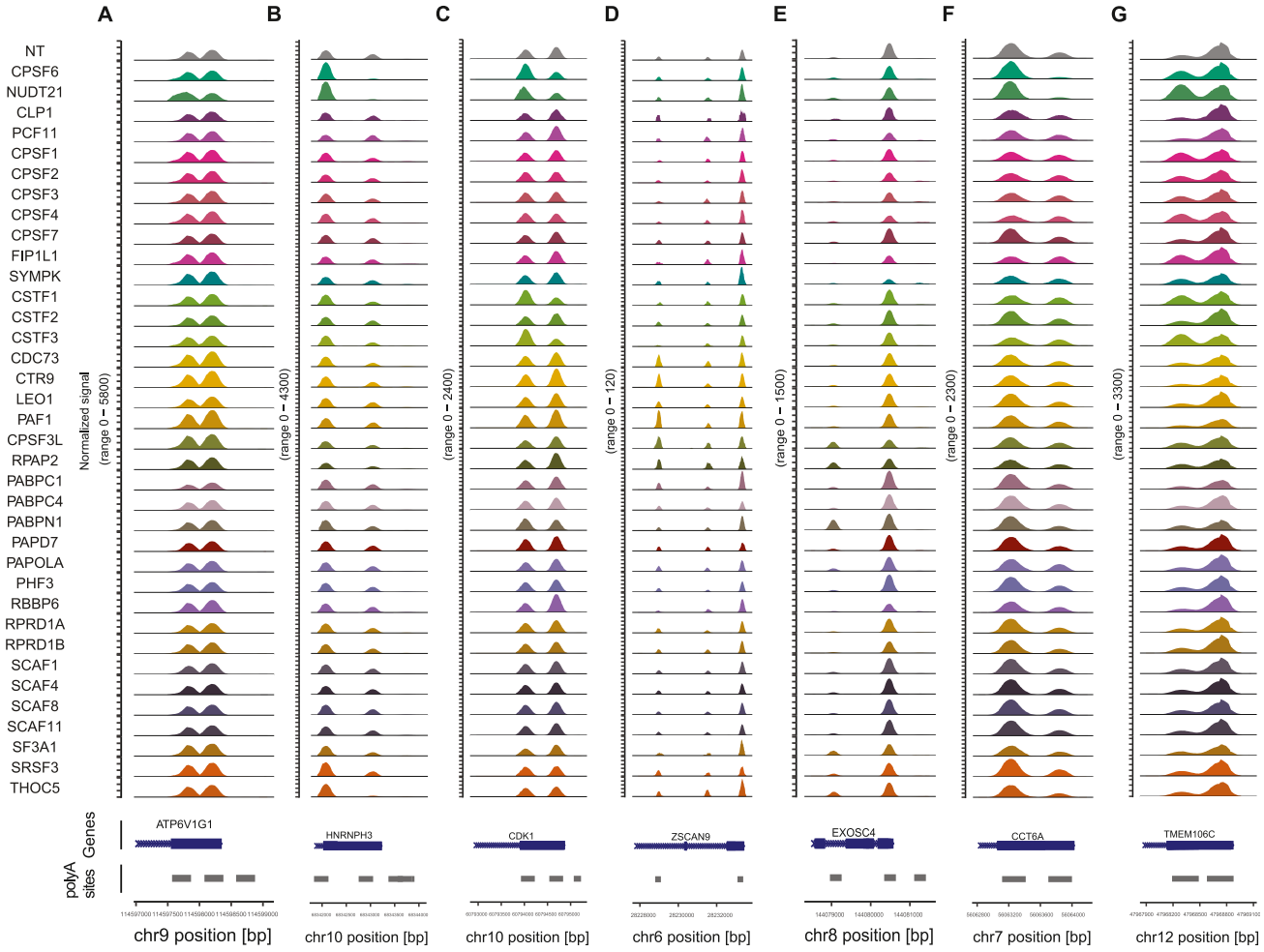
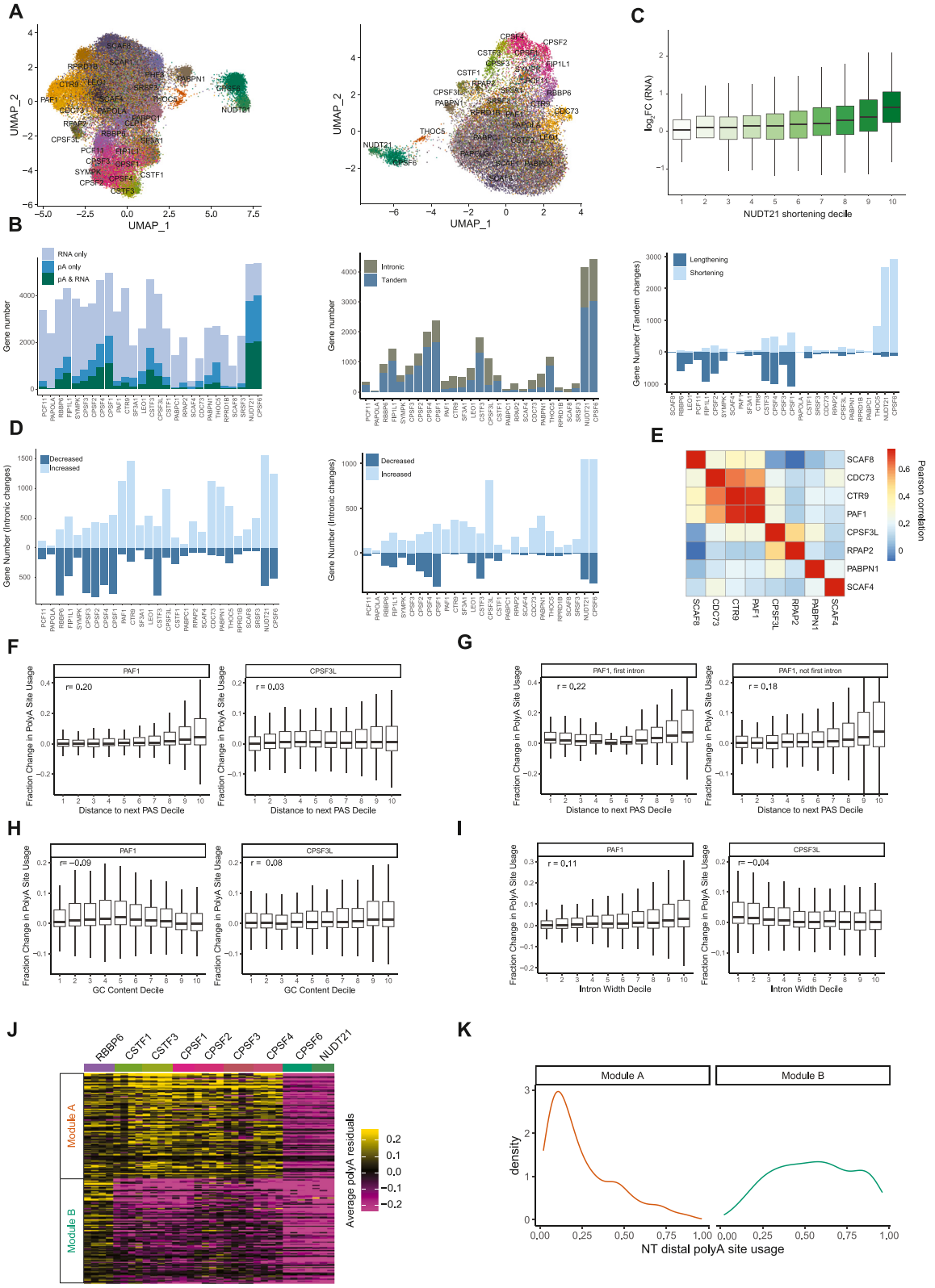


Figure S3. Representative read coverage plots in CPA-Perturb-Seq dataset, related to Figure 1

(A–G) Read coverage plots depicting the differential use of alternative polyA sites at representative loci. Each track represents a pseudobulk average of cells, grouped by their target gene perturbation. Loci were selected based on read coverage plots shown in Figures 1, 2, 3, 4, and 5 and show data for 36 regulators and NT control cells. At the ATP6V1G1 locus (A), we primarily observed changes in total gene abundance, rather than relative polyA site usage, across regulators.



(legend on next page)

Figure S4. Perturbation-driven changes in tandem and intronic polyadenylation, related to Figures 3 and 4

(A) UMAP visualization of HEK293FT (left) and K562 (right) cells profiled via CPA-Perturb-seq. Cells are colored by target gene identity (colors for each gene are the same as those used in Figure 1C). Visualization was computed based on a linear discriminant analysis (LDA) of polyA residuals of tandem and intronic polyA sites.

(B) Same as Figure 3A but from the K562 dataset.

(C) Boxplot indicating the observed $\log_2$ fold change in gene expression after NUDT21 perturbation (HEK293FT cells). Genes are partitioned into deciles based on the degree of 3′ UTR changes observed after NUDT21 perturbation.

(D) Number of genes with changes in intronic polyA site usage, classified by increased or decreased usage of intronic sites for both HEK293FT (left) and K562 (right) cells. Regulators PAF1, CTR9, CDC73, SCAF8, SCAF4, CPSF3L, RABP2, and PABPN1 primarily demonstrate increased usage of intronic polyA sites.

(E) Pearson correlation matrix depicting the relationships between intronic polyadenylation regulators in HEK293FT cells. Correlations are calculated using the linear model coefficients learned during differential polyadenylation analysis (STAR Methods). Genes are ordered via hierarchical clustering.

(F) Percent change (y axis) in polyA site usage of PAF1-perturbed (left) and CPSF3L-perturbed (right) cells compared to NT control cells for intronic polyA sites as a function of distance to the next polyA site (binned into x axis deciles). The distance to the next polyA site correlates with the response of an intronic site to PAF1, but not CPSF3L, perturbation. Pearson correlation is computed between the distance (log-scale) and percent change.

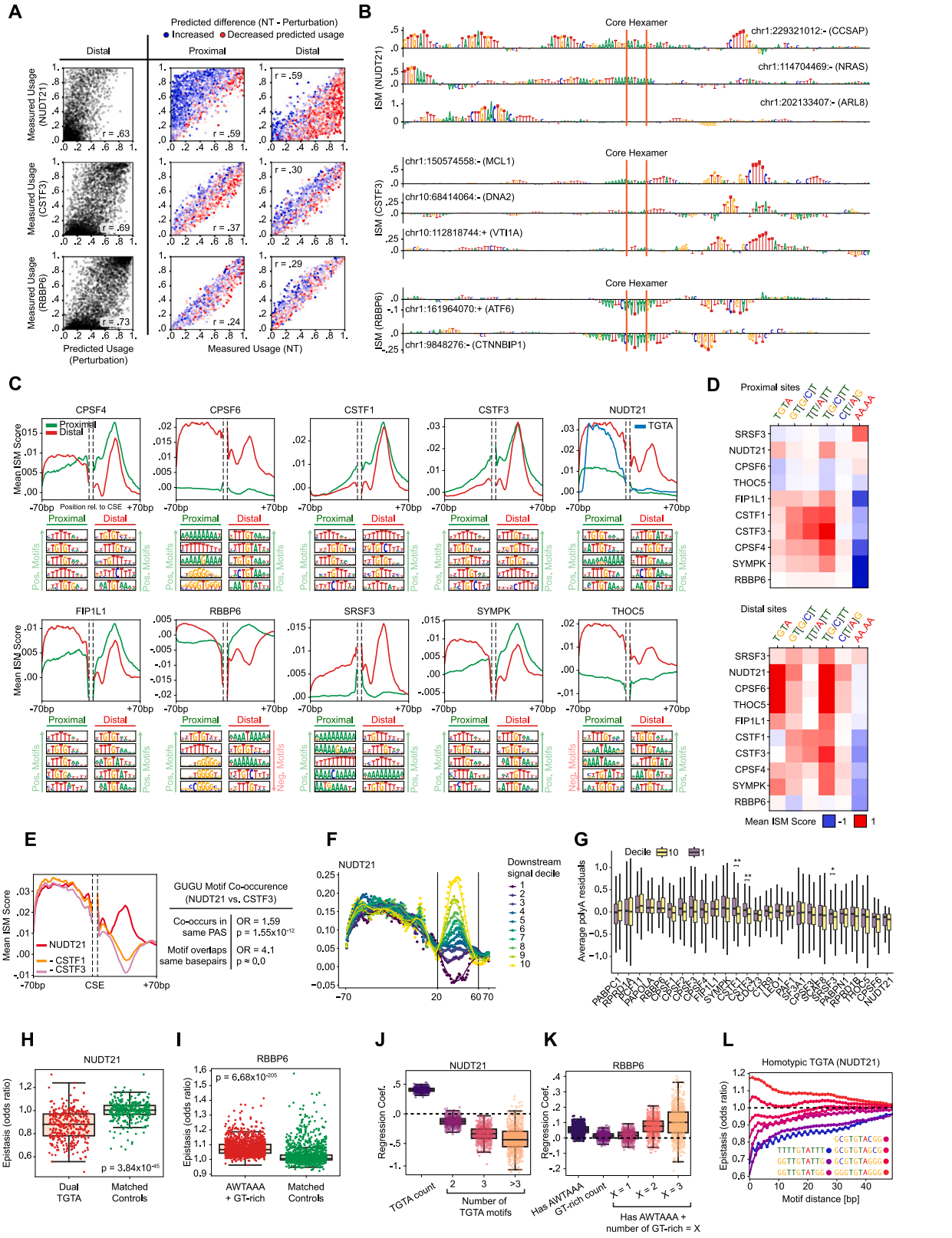
(G) Same as (F) but shown for sites in the first intron (left) and not in the first intron (right) for PAF1 perturbation.

(H) Same as (F) but binning intronic polyA sites into deciles based on GC content.

(I) Same as (F) but binning intronic polyA sites into deciles based on intron width.

(J) Heatmap showing K562 polyA residuals for distal sites in module A genes and module B genes. Identity and order of polyA sites shown in the heatmap are the same as in Figure 4C.

(K) Same as Figure 4F, showing module A and B distal site usage in K562 cells.



(legend on next page)

Figure S5. Evaluation and interpretation of APARENT-Perturb, related to Figure 5

(A) (Left) Predicted vs. measured distal polyA site usage in three perturbation conditions (NUDT21, CSTF3, and RBBP6). (Right) Predicted vs. measured difference in proximal or distal polyA site usage with respect to the non-targeting (NT) condition. Only predictions from the held-out sets of the cross-validation procedure are included.

(B) Same as in Figure 5C but for additional loci.

(C) Average residual attribution scores for all 10 learned perturbations, separated by proximal and distal polyA sites. The corresponding top 5 MoDISco motifs are shown below each plot for proximal and distal sites. These motifs are all associated with positive contribution scores (i.e., they increase perturbation magnitude), except for distal motifs in RBBP6 and proximal motifs in THOC5; these motifs are associated with negative contribution scores (they decrease perturbation magnitude).

(D) Average attribution scores of specific classes of MoDISco motif hits, grouped by perturbation. Red, the motif class is associated with an increase in perturbation magnitude. Blue, the motif class is associated with a decrease in perturbation magnitude.

(E) (Left) Co-occurrence of sequence features with high attribution scores in both the NUDT21 and CSTF1/3 perturbations. Average attribution scores in the NUDT21 perturbation for distal pA signals (red) and average difference in attribution scores between the NUDT21 perturbation and CSTF1 (orange) or CSTF3 (purple). (Right) Co-occurrence analysis of MoDISco motif hits matching U/G-rich motifs in the NUDT21 and CSTF3 perturbations (odds ratios and *p* values calculated with Fisher's exact test).

(F) Same as Figure 5D but showing all distal polyA sites that respond to NUDT21 perturbation. PolyA sites are split into deciles based on the model-assigned importance scores for the downstream element region (DSE; denoted by vertical lines). For decile 1 sites, the DSE is associated with positive attribution scores for NUDT21, while decile 10 sites have negative contribution scores.

(G) Average polyA residuals (scaled by perturbation) for regulators in the CPA-Perturb-seq dataset for polyA sites in decile 1 vs. decile 10. These deciles were defined by the NUDT21 perturbation model but also respond differently to CSTF perturbation. * indicates p value <0.05 in t test comparing residuals in decile 1 vs. decile 10; ** indicates p value <0.001.

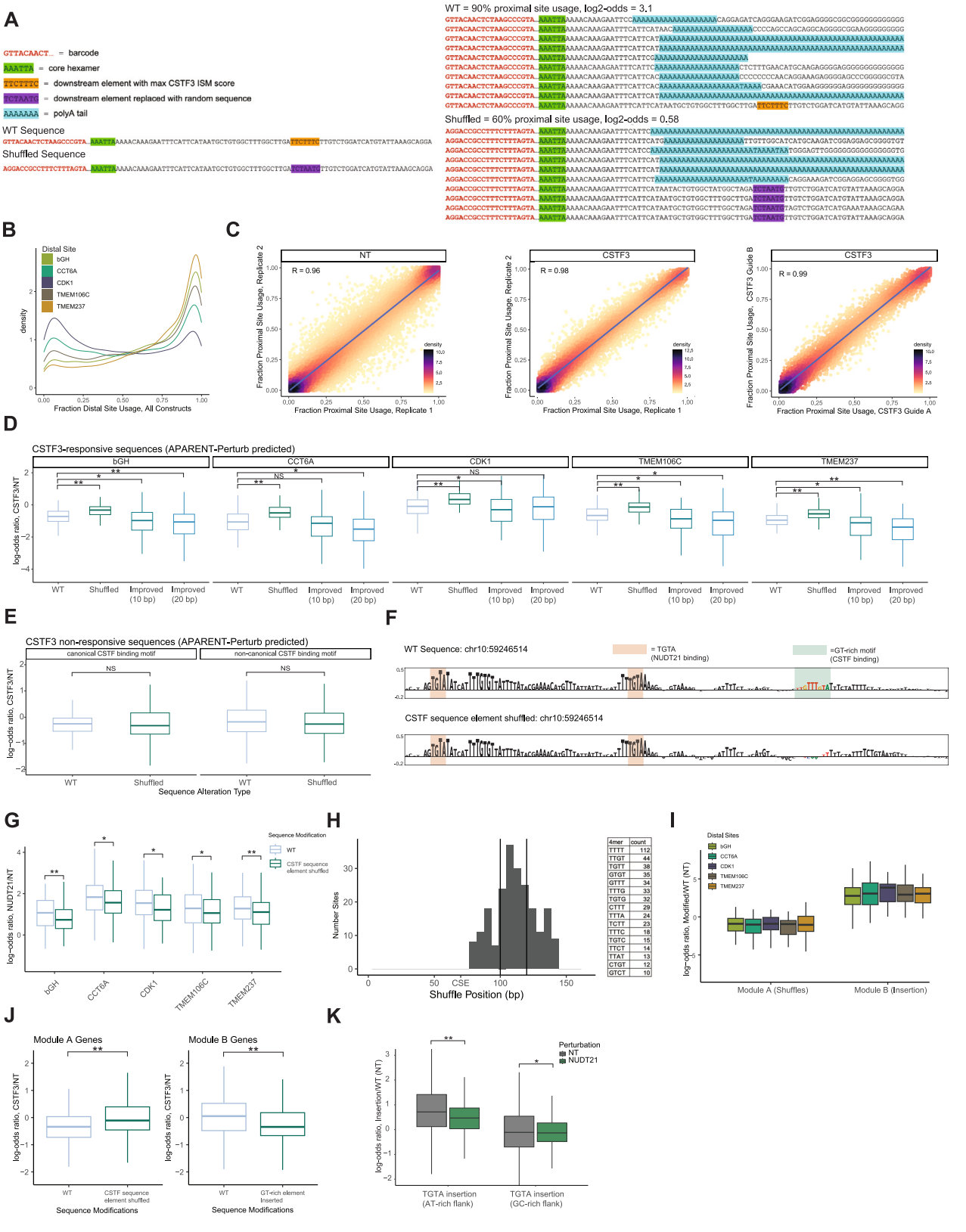
(H) Combinatorial ablation of UGUA motifs in distal polyA signals with exactly two wild-type UGUA motifs in the upstream region, which indicates an overall sub-additive interaction. Epistasis odds ratio (y axis) estimated by exponentiating the difference in predicted perturbation log odds in the NUDT21 condition when replacing both motifs with random sequence and the sum of log odds when replacing one motif with random sequence at a time.

(I) Motif insertion simulation for dual UGUA motifs with varying flanking nucleotide context. The y axis corresponds to the exponentiated difference between the predicted perturbation log odds in the NUDT21 condition when inserting both motifs and when inserting each motif one at a time.

(J) Combinatorial ablation of canonical core hexamer motifs and downstream U/G-rich motifs. The U/G-rich motif is replaced with random sequence, while the core hexamer is replaced with a randomly chosen single-nucleotide mutated hexamer.

(K) Linear regression coefficients of count features and combinatorial indicator variables when fitting the regression model on distal perturbation log-odds ratios in the NUDT21 condition with respect to NT. Distributions of coefficients were generated from 1,000-fold bootstrapping.

(L) Regression coefficient analysis when fitting on RBBP6 perturbation log-odds ratios.



(legend on next page)

Figure S6. Profiling thousands of APARENT-Perturb predictions in multiple genetic contexts, related to Figure 6

(A) Schematic illustrating example constructs of both a WT and shuffled sequence that are inserted into the MPRA construct as well as representative reads for mRNA molecules originating from each sequence. The first 20 bp of the construct consist of a unique barcode (red) that ensures identifiability of each proximal site construct. Each construct is also anchored on a core hexamer (green). In the example shown, we identified a 9-bp downstream element that predicts CSTF3 perturbation response (orange) and replaced that element with random sequence (purple). By directly reading out proximal site usage (when we sequence the polyA tail, highlighted in blue), we measure how sequence modifications change polyA site usage.

(B) Density plot showing fraction distal site usage for all 3,802 tested sequences, grouped into the 5 distal site contexts.

(C) Scatterplot indicating reproducibility of polyA site usage measurements (fraction proximal site usage) across biological replicates for NT (left) and CSTF3 (middle), colored by kernel density estimates. We also observed high reproducibility for independent guides targeting CSTF3 (right).

(D) Same as Figure 6F but separated by distal site context.

(E) Same as Figure 6F but for CSTF3-nonresponsive sequences, as predicted by APARENT-Perturb.

(F) Attribution scores from the APARENT-Perturb NUDT21 model at the distal site of the FAM13C gene (chr10:59246514), both for the wild-type (WT) sequence (top) and upon shuffling the predicted CSTF3 sequence element highlighted in green (bottom). Upstream NUDT21 binding sites (TGTA) are shown in orange.

(G) Same as Figure 6G but split by distal site.

(H) Histogram of the distance (downstream of the core hexamer) for 9-bp regions that were predicted by APARENT-Perturb to have maximal attribution scores for the CSTF3 model (left). Most frequent 4-mers in the WT sequences of these regions include T- and GT-rich elements but not NUDT21 binding sites TGTA (right).

(I) Shuffling the predicted CSTF responsive element in module A genes (n = 49, left) or inserting a GT-rich region in module B genes (n = 47, right) has opposite effects on proximal site usage (log-odds ratio comparing modified sequence to WT, y axis), which is consistent across 5 distal sites.

(J) (Left) CSTF3 perturbation response (log-odds ratio, y axis) for proximal site usage of WT module A genes and upon performing sequence shuffles. (Right) Perturbation response for WT module B sites and upon performing sequence insertions. ** indicates p value <0.0001.

(K) Effect of insertion of TGTA motif with AT-rich (left) or GC-rich (right) flanks on proximal site usage (log-odds ratio of insertion compared to WT sequence, y axis). ** indicates p value <0.0001; * indicates p value = 0.0001–0.05.

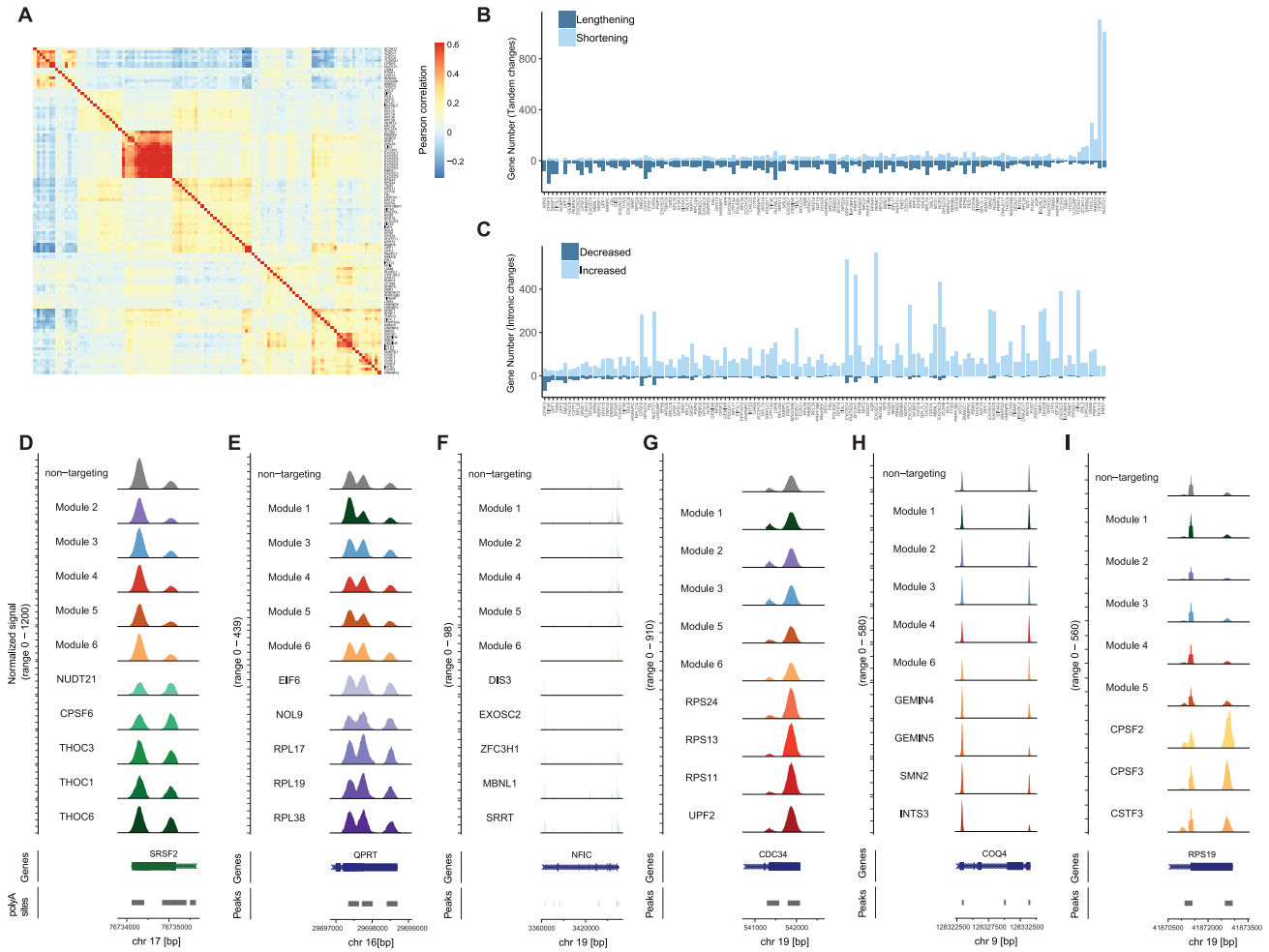


Figure S7. Identifying regulators of alternative polyadenylation from genome-wide Perturb-seq, related to Figure 7

(A) Same as Figure 7C. Axes show all gene names in the correlation matrix.

(B) Number of genes with significant changes in tandem polyA site usage in GWPS dataset, classified by 3′ UTR shortening or 3′ UTR lengthening.

(C) Number of genes with significant changes in intronic polyA site usage in GWPS dataset, classified by increased or decreased intronic polyA site usage.

(D–I) Coverage plots depicting the differential sequencing read coverage and use of alternative polyA sites for representative genes in each module.