Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Dynamics of intronic polyadenylation in the hematopoietic lineage and its regulation by DNA methylation

Richa Rashmi, Abhinaya Muruganandham, Pranita Borkar, et al.

Genome Res. published online April 13, 2026
Access the most recent version at doi:10.1101/gr.281044.125

P<P Published online April 13, 2026 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Dynamics of intronic polyadenylation in the hematopoietic lineage and its regulation by DNA methylation

Richa Rashmi[1], Abhinaya Muruganandham[1,2], Pranita Borkar[1], Sumana Mallick[1,2], Taylor Hubbs[1], Ari Aviles[1,3], Daniel Chung[1], Irtisha Singh[1,2,3*]

[1]Department of Cell Biology & Genetics, Texas A&M University Health Science Center, Bryan, TX, 77807, USA; [2]Department of Biomedical Engineering, Texas A&M University, College Station, TX, 77843, USA; [3]Interdisciplinary Program in Genetics and Genomics, Texas A&M University, College Station, TX, 77843, USA

Abstract

Intronic polyadenylation (IPA) is a key mechanism driving transcriptome diversity, yet its detection and functional characterization remain challenging due to complex splicing patterns and complexity of intronic regions. Here, we introduce IPAseek, a dynamic programming-based computational framework that leverages the Pruned Exact Linear Time (PELT) algorithm and Changepoints Over a Range of PenaltieS (CROPS) to enable *de novo* identification of IPA events from bulk RNA-seq data. IPAseek robustly detects both composite and skipped IPA isoforms.

Applying IPAseek to bulk RNA-seq of hematopoietic cell-types, reveals lineage and stage-specific IPA signatures, with lymphoid cells exhibiting higher IPA site usage than myeloid cells. Temporal profiling during megakaryocyte differentiation uncovers dynamic, gene-specific IPA regulation linked to functional pathways including peroxisomal metabolism and autophagy which are known to play a crucial role in megakaryocytic differentiation, impacting the development and maturation of megakaryocytes. Further, integrative analysis demonstrates that IPA site usage is associated with lower DNA methylation within introns, supporting a regulatory axis connecting epigenetic state and IPA. This finding aligns with emerging evidence that DNA methylation modulates

alternative polyadenylation via CTCF-mediated chromatin looping. Thus, IPAseek provides a platform to characterize IPA across physiological systems and disease contexts using widely available bulk RNA-seq data. These IPA events can be further integrated with other regulatory datasets to elucidate their interplay and functional significance.

Introduction

Cleavage and polyadenylation are a central step in mRNA biogenesis that determines transcript 3′ ends and contributes to gene regulatory complexity. Extensive work over the past two decades has established alternative cleavage and polyadenylation (APA) as a widespread mechanism for generating transcript isoforms with distinct 3′ untranslated regions (3′ UTRs) (Tian et al. 2005; Wang et al. 2008; Proudfoot 2011; Elkon et al. 2013). Variation in 3′ UTR length can influence post-transcriptional regulation by modulating the presence of *cis*-regulatory elements that affect mRNA stability, localization, and translation (Mayr and Bartel 2009; Berkovits and Mayr 2015; Mayr 2019). Consistent with these regulatory roles, APA has been implicated in diverse biological processes, including immune activation and tumorigenesis, while genetic variation near polyadenylation signals (PAS) has been linked to disease-associated traits (Sandberg et al. 2008; Neve et al. 2017; Gabel et al. 2024).

Intronic polyadenylation (IPA) represents a distinct class of APA in which RNA isoform terminates within intronic regions, generating truncated mRNA isoforms with altered coding region (Tian et al. 2007; Singh et al. 2018). IPA events can arise through different splicing configurations, including composite terminal exons or skipped terminal exons, and have been shown to expand transcript diversity. A well-studied example occurs at the immunoglobulin heavy chain locus, where developmentally regulated IPA site usage in B cells controls the production of membrane-bound versus secreted IgM (Early et al. 1980; Tian et al. 2007). More broadly, IPA has been associated with cell differentiation, immune function, stress responses, and cancer, underscoring its potential impact on gene regulation and cellular identity (Ni and Kuperwasser 2016; Singh et al. 2018; Zhao et al. 2021; Sun et al. 2024). Despite these observations, the factors that shape cell-type specific IPA site usage remain largely under explored.

Although numerous computational approaches have been developed to study APA using RNA sequencing data, most were designed to detect 3' cleavage events within annotated 3′ UTRs and are not optimized for intronic regions (Xia et al. 2014; Gruber et al. 2018; Ha et al. 2018). Detecting IPA poses additional challenges due to lower transcript abundance, complex splicing patterns, and confounding signals from intron retention. While experimental methods such as 3′-end sequencing provide high-resolution maps of polyadenylated RNA cleavage sites, their scalability limits their use in large datasets (Derti et al. 2012; Lianoglou et al. 2013). Recent bulk RNA-seq based tools like IPAFinder and InPACT, targeting IPA have begun to address these limitations but remain sensitive to coverage biases or sequence-based assumptions, highlighting the need for robust methods specifically tailored to genome-wide analysis of IPA dynamics (Scotto-Lavino et al. 2006; Liu et al. 2024).

Hematopoietic differentiation provides a well-defined biological framework in which to examine dynamic RNA processing events across closely related cell states. Epigenetic mechanisms, including DNA methylation, play essential roles in regulating gene expression during hematopoiesis and have been implicated in modulating polyadenylation site usage in selective contexts (Wood et al. 2008; Cowley et al. 2012; Nanavaty et al. 2020). However, the extent to which DNA methylation contributes to cell-type specific regulation of IPA remains unclear.

In this study, we investigate the dynamics of IPA across hematopoietic lineages using genome-wide transcriptomic and epigenomic data. We develop and apply a computational framework called IPAseek for systematic identification and quantification of IPA events and use this approach to characterize lineage-associated patterns of IPA site usage. We further examine the relationship between DNA methylation and IPA to explore potential epigenetic contributions to cell-type specific IPA regulation.

Results

IPAseek - A Method for IPA Site Detection

Changepoints are positions within a data sequence where there is a shift in statistical properties, such as mean or variance. In the context of bulk RNA-seq data, a used IPA site manifests as a transition in transcript coverage: there is coverage upstream of the site and a drop or absence of coverage downstream, creating a clear changepoint signature (composite IPA, **Fig. 1A**). This pattern can become more complex when splicing events are coupled with 3′-end formation, resulting in multiple changepoints (e.g., skipped IPA, **Fig. 1D**), and bulk RNA-seq coverage biases further add to the challenges of changepoint detection (Killick et al. 2012; Haynes et al. 2017).

To enable *de novo* detection of IPA sites from individual bulk RNA-seq profiles, we developed a computationally efficient methodology capable of identifying multiple changepoints within the vast search space of intronic regions of expressed genes. Our approach leverages the Pruned Exact Linear Time (PELT) algorithm, a dynamic programming method with linear computational complexity that applies a penalty for each additional changepoint to balance sensitivity and specificity (Killick et al. 2012). Because coverage biases and inherent fluctuations in bulk RNA-seq data can lead to overfitting (too many changepoints) or underfitting (too few changepoints) when using a fixed penalty, we integrated PELT with the Changepoints Over a Range of PenaltieS

(CROPS) method, which systematically explores optimal segmentations across a spectrum of penalty values (Haynes et al. 2017).

Our algorithm proceeds as follows: i) Bulk RNA-seq reads are aligned to the genome to generate coverage profiles; ii) introns of protein-coding genes that exhibit retention are identified and filtered out to avoid spurious signals; iii) introns of protein-coding genes with contiguous coverage over at least 100 consecutive bases are selected, as very short covered regions within introns might not be functionally relevant; iv) PELT is applied to these coverage profiles to detect changepoints; and v) differential expression upstream and downstream of each detected changepoint is assessed to confirm a shift in the coverage profile (see Methods and **Supplementary Fig. S1**). By running changepoint detection over a range of penalties, the algorithm determines the optimal number of changepoints and evaluates the risk of overfitting at lower penalties and underfitting at higher penalties (**Fig. 1B-C, 1E-F**). The resulting changepoints are further evaluated by checking for junction-spanning reads upstream and downstream to distinguish IPA sites from splicing events.

A key feature of IPAseek is its use of per-base resolution coverage data from uniquely mapping reads across introns and flanking coding regions, providing a high-resolution foundation for changepoint detection. The integration of the PELT algorithm with CROPS, combined with filtering for intron retention and splicing, enables efficient and robust identification of both single and multiple changepoints in bulk RNA-seq datasets, thereby enhancing the accuracy and reliability of IPA site detection (**Fig. 1**).

Enhanced Detection of IPA Sites by IPAseek Using Bulk RNA Expression Coverage

To evaluate IPAseek performance in *de novo* IPA detection, we applied it to 18 test bulk RNA-seq samples that also had paired 3′-seq data (see **Supplementary_Table_S1**) (Lee et al. 2018;

Samur et al. 2018). We first assessed the quality of these bulk RNA-seq samples by comparing reads aligned to intronic versus exonic regions (**Supplementary Fig. S2A**), and categorized them as high, moderate, or low-quality based on the relative coverage in these regions.

To define ground truth for IPA site validation, we used IPA sites detected with ≥ 5 TPM (tags per million) expression in the 3′-seq atlas as described by Singh et al. (Lee et al. 2018; Singh et al. 2018). Currently, 3′ end sequencing techniques represent the state-of-the-art methodology for the precise identification and quantification of transcript 3′ ends. Thus, after removal of known artifacts as described by Singh et al., it was used as a reference standard to evaluate the performance of IPAseek. Because achieving single-nucleotide precision from bulk RNA-seq is challenging due to read length and coverage resolution (Shenker et al. 2015; Arefeen et al. 2018), IPA sites detected within 350 nt of a ground-truth site were considered true positives (TP), ground-truth sites missed in RNA-seq were considered false negatives (FN), and IPA sites detected only in RNA-seq or expressed at ≤ 5 TPM in the 3′-seq atlas were categorized as false positives (FP), hereafter referred to as uncharacterized IPA sites.

Using this framework, IPAseek identified IPA events that were not detected by IPAFinder or InPACT but were supported by ground truth. These included a composite IPA site in *ING5* (106 bp upstream of the 3′-seq sites; **Fig. 2A**) and a skipped IPA site in *EXOC4* (185 bp upstream; **Fig. 2B**). We validated the 3′-end of the *EXOC4* IPA isoform using 3′-RACE (**Fig. 2C**), confirming that IPAseek detected a cleavage site 184 bp upstream of the site mapped by 3′-RACE.

To further evaluate the biological relevance of IPAseek-detected events, we analyzed PAS enrichment within ±200 nt of IPA sites and observed enrichment for canonical AAUAAA/AUUAAA motifs across all test samples (**Supplementary Fig. S2B**), consistent with known PAS driven cleavage (Tian et al. 2005). Approximately 37% of IPAseek-detected IPA isoforms retained ≤25% of the coding sequence (**Supplementary Fig. S2C**), aligning with previous reports that many IPA isoforms truncate early in the transcription unit (Singh et al. 2018). Together, these results demonstrate that IPAseek can detect biologically relevant IPA events.

Enhanced IPA Detection with IPAseek

We next evaluated IPAseek's performance more systematically by comparing it with two published IPA detection methods, InPACT and IPAFinder, using precision, recall, and F1-score metrics (**Fig. 3A**, **Supplementary Fig. S3A**). Across all 18 test samples, IPAseek achieved a mean precision of 0.16, mean recall of 0.08, and mean F1-score of 0.10, compared with 0.12, 0.06, and 0.08, respectively, for IPAFinder and 0.05, 0.03, and 0.03 for InPACT (see **Supplementary_Table_S2**). When stratified by sample quality, IPAseek demonstrated improvement in metrics within each group; in high-quality samples, precision, recall, and F1-score reached 0.28, 0.14, and 0.19, respectively, outperforming IPAFinder and InPACT.

As demonstrated above, under the default settings, IPAseek performs best on high-quality RNA, with more modest gains over IPAFinder in moderate and low-quality samples. To improve sensitivity, we relaxed the intron retention filter, defined as the ratio of median intronic to upstream exon coverage, which increased detectable IPA events but also raised false positives, in a sample-quality dependent manner. These user-defined thresholds allow IPAseek to be tuned to different data qualities and tolerances for false positives.

Even though IPAseek outperforms the other methods, the mean recall remains low, indicating that a substantial number of IPAs are hard to detect from bulk RNA-seq alone. To examine factors influencing detection, we compared terminal IPA exon expression between true positives and false negatives and found that true positives exhibit significantly higher expression than false negatives across multiple test samples (**Fig. 3B**, **Supplementary Fig. S2D**). This supports IPA expression level to be a key determinant of detectability in bulk RNA-seq based IPA calling.

To investigate the validity of false positives, we assessed whether these sites represent authentic IPA events that are unannotated or missed by our 3′-seq cutoff or instead reflect artifacts. We examined additional evidence for 3′-end formation by querying PolyASite, RefSeq, Ensembl, and lowly expressed 3′-seq events (≤ 5 TPM), and by assessing the presence of essential PAS motifs near the determined cleavage sites (O'Leary et al. 2016; Herrmann et al. 2020; Martin et al. 2023). IPAseek detected fewer false positives overall, especially in high-quality samples, than IPAFinder and InPACT (**Fig. 3C**, **Supplementary Fig. S3B**). For example, in Test Samples 1, 2, 3, 4, 8, and 11, IPAseek identified 466, 474, 547, 473, 459, and 377 false positives, respectively, compared with 727, 460, 731, 594, 508, and 547 for IPAFinder and 315, 160, 1874, 1469, 1275, and 1742 for InPACT.

Most IPAseek false positives could be cross-referenced to existing databases or supported by lowly expressed isoforms in the 3′-seq atlas, indicating that many of these sites likely correspond to bonafide IPA events not captured by the stringent ground-truth definition. False positives from InPACT and IPAFinder also contained previously unannotated sites, underscoring the potential of all three methods to expand the landscape of known IPA events. To further evaluate the biological relevance of unannotated IPAseek false positives, we examined PAS motifs within ±150 nt of the potential cleavage sites and observed a strong enrichment for canonical and non-canonical PAS sequences, with most motifs located close to the IPA sites (**Fig. 3D-E, Supplementary Fig. S3C-D**).

We conducted further characterization of selected false positive sites and identified a distinct composite IPA event within the *PARVB* gene (Chr22:44024452-44024923 (+)) and a skipped IPA event in the *CEACAM8* gene (Chr19:42587622-42588409 (−)) (**Fig. 3F-G**). These IPA events were not detected by the two other methods, InPACT and IPAFinder. Collectively, these results establish IPAseek as an effective tool for detecting diverse IPA events across complex transcriptomic landscapes. Overall, IPAseek represents an advancement in transcriptomics research with implications for understanding gene expression regulation in both physiological and disease contexts.

IPAseek Reveals Stage-Specific IPA Site Usage During Megakaryocyte Differentiation

In this section, we used IPAseek to systematically analyze changes in IPA during the differentiation of myeloid progenitor K562 cells into megakaryocytes. Bulk RNA-seq data from eight sequential time points spanning 0-4320 min, with two biological replicates per time point, were obtained from the GSE213909 dataset (Bond et al. 2023). IPAseek identified 2,342 confident IPA events across replicates (present in both replicates at any time point), with 225 highly used events (≥ 25% usage in ≥ 2 samples) selected for detailed analysis.

Temporal analysis contrasting early (0-360 min) and late (1440-4320 min) stages revealed distinct clusters of IPA events with differential usage (**Fig. 4A**). A heatmap visualizing dynamic IPA site usage showed clear separation of early and late time points, demonstrating progressive rewiring of the IPA landscape over 72 hours (**Fig. 4A**).

Our analysis highlighted two genes, *ACOT8* and *PEX13*, exhibiting contrasting patterns of IPA site usage linked to differentiation stage (**Fig. 4A**). *ACOT8* showed increased IPA site recognition at later time points (**Fig. 4B, 4F, Supplementary Fig. S4B**), while *PEX13* displayed reduction in IPA site usage as differentiation proceeded (**Fig. 4C, 4G, Supplementary Fig. S4C**). Protein-

protein interaction network analysis confirmed a direct functional relationship between *ACOT8* and *PEX13* (**Supplementary Fig. S4A**), supporting coordinated regulation of peroxisomal function and lipid metabolism during megakaryopoiesis (Hunt et al. 2012; Lee et al. 2018; Plessner et al. 2024).

To validate these IPA events, we performed 3′-RACE for *ACOT8* and *PEX13*, confirming the IPA isoforms detected by IPAseek (**Fig. 4D**). 3′-RACE revealed an additional intronic cleavage site in *PEX13* within the intronic TE, suggesting multiple IPA cleavage sites around this locus. However, the complexity of the locus and bulk RNA-seq resolution constrained IPAseek's ability to reliably detect the additional site (detected at multiple time-points but not consistently across replicates).

Megakaryocytic differentiation was induced in K562 cells using 25 nM PMA and confirmed by phase-contrast microscopy showing hallmark morphological changes (**Fig. 4E**) and upregulation of *ITGB3* (**Supplementary Fig. S4D**). In-house time-course validation showed *ACOT8* IPA site usage increasing and *PEX13* IPA site usage decreasing over time (**Fig. 4F, G**), mirroring bulk RNA-seq patterns determined from the datasets obtained from the public domain (**Supplementary Fig. S4E-F**).

Together, these results demonstrate that IPAseek captures stage-specific IPA events during megakaryocyte differentiation.

Comprehensive Atlas of IPA events Reveals Lineage and Stage Specific Regulation in Hematopoietic Lineage Cells

To systematically characterize IPA events across hematopoietic cell-types, we constructed an atlas using bulk RNA-seq data from 78 samples representing 10 hematopoietic populations, including differentiated cells (myeloid: granulocytes, monocytes; lymphoid: naïve B cells, $CD4^+$/$CD8^+$ T cells, NK cells) and undifferentiated precursors (granulocytic, monocytic, erythroid precursors, and $CD34^+$ HSPCs).

To assess whether IPA site usage distinguishes cell lineages and differentiation states, PCA followed by *k*-means clustering (k = 3) of IPA site usage in differentiated cells (GSE184264) revealed three distinct clusters: granulocytes (myeloid; cluster 1, n = 7, silhouette width = 0.36), lymphoid/monocytes (clusters 2/3, n = 13/19, silhouette widths =0.52/0.56), with overall average silhouette width of 0.511 (**Fig. 5A**, **Supplementary Fig. S5C**). To further enhance the separation, we repeated the PCA and clustering excluding granulocytes (**Supplementary Fig. S5A**). In this subset, k-means (k=5) identified five clusters, with most monocyte samples (5 out of 7) forming Cluster 1, and all lymphoid samples distributed among the remaining clusters. For undifferentiated precursors (GSE114922), *k*-means clustering (k = 4) identified four groups with HSPCs dominating clusters 1/2 (n = 65/17, silhouette widths = 0.49/0.71), overall average silhouette width 0.52 (**Fig. 5B**, **Supplementary Fig. S5D**). Given the higher number of HSPC samples, we performed PCA and clustering after excluding HSPCs (**Supplementary Fig. S5B**). This analysis identified three groups, with granulocytic precursors and monocytic precursors primarily clustering together (Cluster 1), demonstrating their lineage similarity. We did this independently for the two datasets to avoid factors like differences in sample preparation, sequencing techniques etc. that could convolute the interpretation of results.

To identify lineage-enriched IPA events, differential analysis between myeloid and lymphoid lineages identified 981 lineage-enriched IPA events (Wilcoxon rank-sum test, $P-\mathrm{adj}$ < 0.05), with hierarchical clustering revealing clear separation (**Fig. 5C**). Similarly, analysis of 36 undifferentiated precursor samples identified 372 cell-type specific IPA events (Kruskal-Wallis test, $P-\mathrm{adj} < 0.05$), showing distinct patterns among granulocytic precursors, monocytic precursors, and CD34$^+$ HSPCs (**Fig. 5D**).

RPN1 exemplified lymphoid-enriched IPA site usage (**Fig. 5C, E**), with bulk RNA-seq coverage confirming ~4.5-fold higher IPA site usage in lymphoid versus myeloid cells (Wilcoxon rank-sum test, $P-\mathrm{adj} < 0.05$). Together, these validated clusters and differential analyses demonstrate lineage and stage-specific IPA regulation during hematopoiesis.

Distinct DNA Methylation Landscapes Surround IPA Sites in Hematopoietic Lineage Cells

Although IPA is widespread, gene and cell-type specific, and responsive to stimuli, we know little about factors enhancing or preventing its usage (Lee et al. 2018; Singh et al. 2018). Epigenetic modifications like DNA methylation modulate gene expression without altering DNA sequence, which is analogous to regulated IPA site recognition, though IPA differs from 3' UTR APA due to concurrent splicing.

Prior studies showed DNA methylation ablation increases proximal PAS usage by disrupting CTCF/cohesin binding (Nanavaty et al. 2020; Fink et al. 2025), and influences allele-specific IPA of imprinted genes H13/Nap1l5 (Wood et al. 2008; Cowley et al. 2012). No studies have investigated DNA modifications regulating cell-type specific IPA site usage in its endogenous state. We therefore examined how endogenous DNA methylation state impacts IPA site selection across immune cells and cancers.

To test this systematically, we integrated 198 bulk RNA-seq and 189 BS-seq samples (RRBS/WGBS) across 12 immune cell-types (**Supplementary Fig. S6A-B**). IPAseek identified

IPA sites from RNA-seq data. We compiled IPA signals (PAS) and categorized them as "used" (within 200 bp of high-confidence IPA: RPKM ≥ 0.5, usage ≥ 10% in ≥10% samples) or "unused" to control for sequence biases (**Supplementary Fig. S6C**).

Among 910 used and 1,820 unused PAS (>5.7M methylation sites), used PAS consistently exhibited higher surrounding DNA methylation (**Fig. 6A**, KS-test, *P*-value $\leq 9.88\times10^{-270}$), indicating methylation influences PAS recognition. This pattern persisted across cell-types: B cells (679 vs. 1358, *P*-value $\leq 5.14\times10^{-177}$; **Fig. 6B**), CD14^{+} monocytes (325 vs. 650, *P*-value $\leq 3.94\times10^{-244}$); **Fig. 6C**), AML cells (552 vs. 1104, *P*-value $< 10^{-244}$; **Fig. 6D**), each analyzing 2.95-3.4M sites.

To test if methylation differences extend intron-wide, we performed rolling window (100 bp windows, 50 bp steps) and equal-tile (500 bins) analyses comparing IPA introns (n = 696) to non-IPA introns from IPA genes (n = 1,758) and non-IPA genes (n = 1,303) (**Fig. 6E**, **Supplementary Fig. S6D**; KS-test *P*-value $\leq 2\times10^{-16}$). Higher methylation across IPA introns suggests broader chromatin context influences cleavage.

To determine if methylation changes drive IPA site usage changes, we analyzed GSE184314 (CD4^{+} cells), where hypomethylated introns (n = 80) were enriched for increased IPA site usage (Fisher's exact, *P*-value $\leq 2.78\times10^{-3}$; **Fig. 6G**) while hypermethylated introns (n = 135) showed no association (**Fig. 6F-G**). Reciprocally, low-usage IPA introns (n = 19) were enriched for increased methylation (*P*-value $\leq 5.24\times10^{-3}$; **Fig. 6H**), but high-usage introns (n = 34) showed no association (**Fig. 6H-I**). Patterns were partially replicated in GSE66117 (**Supplementary Fig. S7A**) but inconsistent elsewhere (**Supplementary Fig. S7B-F**).

Collectively, increased intronic DNA methylation correlates with reduced IPA site usage across contexts, though causality remains untested. Unlike distal CpG island effects on APA, we found robust proximal (≤ 2,500 bp) methylation-IPA associations (**Fig. 6A-E**).

Discussion

In this study, we present IPAseek, a computational framework that leverages dynamic programming and changepoint detection algorithms to accurately identify intronic polyadenylation (IPA) events from bulk RNA-seq data. Unlike previous methods primarily focused on 3' UTR alternative polyadenylation (APA), IPAseek robustly detects both composite and skipped IPA isoforms, addressing challenges posed by complex splicing patterns and complexity in intronic regions. Our benchmarking against paired 3′-end sequencing data and comparison with established tools such as IPAFinder and InPACT demonstrate that IPAseek achieves superior sensitivity and precision, enabling the discovery of IPA events which were also validated by orthogonal techniques like 3'-RACE. This advancement fills a critical in transcriptome analysis, as IPA has been historically under-characterized despite its emerging importance in gene regulation (Gruber et al. 2018; Singh et al. 2018).

Applying IPAseek to a comprehensive dataset of immune cell-types and hematopoietic differentiation time courses, we reveal that IPA site usage is highly dynamic and exhibits lineage and stage-specific patterns. Consistent with previous reports indicating that IPA contributes to transcriptome diversification in hematopoietic cells, we observed that lymphoid cells exhibit higher IPA site usage than myeloid cells (Singh et al. 2018). This finding aligns with the notion that IPA modulates immune cell identity and function, as exemplified by the well-characterized IPA of the immunoglobulin M heavy chain (*IGHM*) locus, which controls the production of secreted versus membrane-bound IgM isoforms in B cells (Early et al. 1980; Takagaki and Manley 1998). Our temporal profiling during megakaryocyte differentiation further highlights gene-specific IPA regulation linked to key biological process like autophagy, extending prior observations that IPA influences gene expression programs during differentiation and stress responses (Thomas et al. 2012; Cheng et al. 2020).

A major contribution of our study is the integrative analysis of endogenous DNA methylation state and IPA site usage, which uncovers a previously underappreciated epigenetic layer modulating IPA. We demonstrate that introns containing IPA sites exhibit elevated DNA methylation compared to non-IPA introns, a pattern that holds true across multiple cell-types. This observation is in line with recent findings that DNA methylation can influence APA by modulating chromatin architecture and the recruitment of RNA processing factors (Cowley et al. 2012; Smith 2019; Nanavaty et al. 2020; Jia et al. 2024; Fink et al. 2025). For instance, DNA methylation-dependent CTCF binding has been shown to regulate APA site choice by altering chromatin looping and polymerase pausing, thereby affecting transcript isoform diversity. Our results extend these insights by demonstrating a genome-wide association between methylation and IPA site usage, supported by enrichment in hypermethylated introns exhibiting reduced IPA site usage.

The complex relationship between methylation and IPA site usage we observed suggests that DNA methylation acts as a context-dependent modulator rather than a simple on/off switch. We observed that hypermethylation is associated with suppression of IPA. This finding underscores the need for further mechanistic studies to dissect how epigenetic states integrate with RNA processing machinery to fine-tune transcript isoform expression in a cell-type and condition-specific manner.

The biological implications of our findings are profound. IPA-mediated transcript truncation can generate non-coding RNAs or protein isoforms with altered functional domains, impacting processes such as immune signaling, cell fate determination, and disease progression. Dysregulated IPA has been implicated in cancer, where IPA events can produce truncated tumor suppressors or oncogenic isoforms, contributing to tumorigenesis and therapy resistance (Mayr and Bartel 2009; Ni and Kuperwasser 2016; Lee et al. 2018; Li et al. 2020; Zhao et al. 2021; Cheng et al. 2024). Our demonstration that DNA methylation is associated with IPA site selection suggests an epigenetic mechanism by which cancer cells and immune cells may modulate transcriptome complexity to adapt to environmental cues or evade immune surveillance.

IPAseek's scalability and accuracy make it a valuable tool for future investigations into the regulatory landscape of IPA across diverse biological contexts. Integrating IPA detection with epigenomic profiling will be important to elucidate how IPA contributes to development and disease. Future studies leveraging single-cell and long-read sequencing technologies may further resolve IPA isoform heterogeneity and its functional consequences at higher resolution. Additionally, experimental perturbation of DNA methylation and chromatin modifiers will be essential to establish causal links between epigenetic states and IPA regulation.

In conclusion, our work establishes IPAseek as a platform for dissecting the complexity of IPA and its epigenetic regulation. By revealing the dynamic interplay between DNA methylation and IPA site usage, we provide insights into the multi-layered control of gene expression.

Methods

IPAseek algorithm

Intron Preprocessing

We obtained the human genome annotation (version hg38) from UCSC RefSeq (table name: refGene) and flattened the entire genome, annotating each position with a genomic region: intron, CDS, 3' UTR, 5' UTR, ncRNA, or intergenic, according to the RefSeq annotation. Upstream/downstream regulatory regions were defined as 5 kb extensions from 5' UTR (5' UTR*) and 3' UTR (3' UTR*) to capture potential regulatory elements influencing the gene expression. When annotating overlapping regions, we applied the following priority order: CDS > 5' UTR > intron > 3' UTR and CDS+3' UTR. Intronic regions were filtered using:

$$L_{intron} \in [500]\ [150000] \text{ bp}$$

where L_{intron} represents intron length. The lower bound (500 bp) ensures sufficient resolution for bulk RNA-seq coverage analysis, while the upper limit (150 kbp) excludes ultra-long introns prone to alignment artifacts. Genes lacking coding regions (CDS=∅) and introns overlapping snoRNA/miRNA loci (UCSC Table Browser, assembly hg38), blacklisted regions (ENCODE), or retrotransposons (RepeatMasker) were excluded to focus on protein-coding transcripts and minimize confounding signals(Griffiths-Jones 2004; Weber 2005; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008).

Sample Preprocessing

For each sample, reads/FASTQ files were aligned to the reference genome using STAR (v2.7.10a) with default parameters (Dobin et al. 2013). To reduce false-positive coverage signals, only uniquely mapped reads (MAPQ ≥ 255) were retained. Gene expression quantification used RPKM normalization:

$$RPKM = \frac{No.of\ reads\ mapped\ to\ the\ gene \times\ 10^9}{Total\ Library\ Size \times\ Gene\ Length\ (bp)}$$

Gene expression levels were quantified using RPKM to account for transcript length and sequencing depth biases (Mortazavi et al. 2008). Genes with RPKM > 0.5 were considered expressed, a conservative threshold balancing sensitivity and specificity in bulk RNA-Seq datasets for IPA site detection.

Retained Intron Identification and Filtering

The IPA detection begins with calculating per-base read coverage from uniquely mapping reads over filtered introns derived from expressed genes. This step also includes calculating coverage for the coding sequence (CDS) flanking the introns under consideration. Following the coverage calculations, we focus on removing retained introns, as these are unlikely to contain premature cleavage sites(Monteuuis et al. 2019). We established four specific criteria for classifying an intron as retained: (1) a minimum of three reads spanning the intron-exon junction, (2) at least 85% of the intron covered by reads, (3) median coverage over the flanking exons exceeding 0.5 RPKM, and (4) a ratio of median coverage over the intron to median coverage over the upstream exon of at least 5% (Middleton et al. 2017). Any intron meeting all four conditions was classified as a retained intron and excluded from further analysis. For the remaining introns, we ensured that they have read coverage to contain a potential TE by checking for a read coverage of at least five over a contiguous stretch of 100 bp. Introns failing to meet this criterion were also excluded from analysis. The next critical step involves detecting significant changepoints in read coverage within the introns, utilizing the Pruned Exact Linear Time (PELT) algorithm.

Changepoint Detection in Introns using PELT

The Pruned Exact Linear Time (PELT) algorithm is an efficient and exact method for detecting multiple changepoints in time series data. It minimizes a penalized cost function that balances segmentation fit with model complexity, defined as:

$$F(n) = \left[\sum_{i=1}^{k+1} C\left(y_{(\tau i-1+1):\tau i} \right) + \beta k \right]$$

where C is the cost function for each segment, k is the number of changepoints, and β is a penalty term that prevents overfitting. The algorithm employs dynamic programming to recursively calculate the optimal segmentation up to each time point and integrates a pruning step to discard suboptimal changepoint candidates. The pruning condition ensures computational efficiency by eliminating candidates that cannot improve the segmentation, reducing the average computational complexity to $O(n)$ under certain conditions.

For bulk RNA-seq analysis, PELT is particularly advantageous due to its ability to adaptively detect abrupt changes in coverage patterns, which may indicate biologically meaningful events such as IPA sites. By adjusting parameters like the penalty value

$$\beta = 2plogn$$

where p is the number of variables and n is the sample size and minimum segment length, detection sensitivity can be fine-tuned based on dataset characteristics. In this study, we applied PELT combined with the Changepoints for a Range Of PenaltieS (CROPS) approach to identify significant changes in bulk RNA-seq intron coverage. CROPS evaluates multiple penalty values within a specified range (100 to 10,000 in our case), enabling adaptive sensitivity in changepoint detection. The implementation was carried out using the cpt.mean function from the "changepoint (version 2.2.1)" R package, which detects mean shifts in time series data. To prepare the data, intron coverage was extracted from BAM files and filtered to retain regions between 500 and 150,000 base pairs in length. This filtering step reduced computational complexity while focusing on biologically relevant regions. For antisense strand introns, coverage vectors were reversed to ensure consistent orientation during processing. The algorithm starts by defining a cost function that measures how well a segmentation fits the data while penalizing each changepoint to prevent overfitting. Using dynamic programming, it recursively calculates the optimal segmentation up to each change point (minimum segment length = 200bp) and applies a pruning rule that eliminates candidate changepoint locations that cannot be part of the optimal solution. This pruning mechanism significantly reduces computational costs without compromising accuracy.

By combining PELT with CROPS, we detected abrupt changes in expression coverage patterns within introns that may correspond to IPA events. The adaptability of this approach allowed us to optimize detection sensitivity while maintaining computational efficiency critical for genome-scale analyses. This method demonstrated robust performance in identifying potential IPA sites while efficiently handling large datasets.

Changepoint Selection

Next, we assessed whether the detected changepoints met IPA conditions. This was achieved by calculating the number of spliced reads within ±50 bp of each changepoint and evaluating RPKM coverage within ±200 bp. For cases where a single changepoint was identified within an intron, we checked for Composite IPA conditions, requiring that the median coverage upstream of the changepoint exceeded the median coverage downstream, with no splice reads present within ±50 bp of the changepoint. When multiple changepoints were detected within an intron, adjacent changepoints were analyzed iteratively in pairs to distinguish between Composite IPA and Skipped IPA scenarios. For Composite IPA detection using changepoint pairs, two conditions had to be satisfied: (1) the median coverage upstream of changepoint 1 must exceed the median coverage downstream of changepoint 1, and (2) the median coverage upstream of changepoint 2 must exceed the median coverage downstream of changepoint 2. Additionally, no splice reads could be present within ±50 bp of either changepoint. For Skipped IPA scenarios, different criteria were applied: (1) the median coverage upstream of changepoint 1 must be lower than the median coverage downstream of changepoint 1, while (2) the median coverage upstream of changepoint 2 must exceed the median coverage downstream of changepoint 2. Furthermore, changepoint 1 required more than 5% spliced reads within ±50 bp, whereas no splice reads could be present within ±50 bp of changepoint 2. (**Supplementary Fig. S1**).

Once potential IPA sites are identified, we annotate the new Terminal Exon (TE) associated with each IPA event. For composite IPA events, the end of the upstream coding sequence (CDS) is designated as the start site of the TE, while the IPA site serves as the end site. For skipped IPA events, the start of the new TE is defined by changepoint 1, and its end is marked by changepoint 2. This approach ensures accurate annotation of TEs for both types of IPA scenarios.

Additional Filtering for Significant IPA Sites

To enhance significance, we applied few additional filters: first, ensuring that coverage (RPKM) 200 bp upstream of each IPA site was greater than 0.5; second, performing differential expression analysis between three contiguous 100 bp windows upstream and downstream of each changepoint using DESeq2 (version 1.46.0) (Love et al. 2014). Only those IPA sites with $P.adj \leq 0.2$ and $P - \mathrm{value} \leq 0.1$ were deemed significant for downstream analysis. Additionally, we filtered out new TE for ≥0.5 RPKM expression. Sites passing these stringent conditions were classified as valid IPA sites along with their respective annotated TEs designated as TEs for the corresponding IPA isoforms.

Construction of an IPA Atlas and Quantification of IPA Site Usage Across Samples

IPAseek can be utilized to create an atlas of IPA events across multiple samples. Due to the resolution limitations of bulk RNA-seq, the same IPA site may be detected a few nucleotides apart in different samples. To address this redundancy when combining IPA events from multiple samples, we merge IPA events that share the same classification (e.g., composite or skipped) and have TE ends annotated within 100 bp of each other. For these merged IPA events, the TE end is assigned as the median of the TE ends from the individual IPA events being combined.

Once the atlas of IPA events is constructed, we assign a binary confidence value to each event. IPA events detected in two or more samples are classified as "confident," while those occurring in only a single sample are labeled as "not confident."

After generating the atlas, we quantify the usage of each IPA event in individual samples. IPAseek calculates the relative usage of an IPA isoform by comparing the expression of its TE to the combined expression of the gene's last coding exon and its TE. This metric, referred to as IPA site usage, provides a normalized measure of how frequently an IPA isoform is utilized relative to other isoforms. The formula for IPA site usage is as follows:

$$IPA\,usage\; = \frac{RPKM_{TE}^{j}}{RPKM_{TE}^{j} + RPKM_{last_{CDS}}^{j}}$$

Here $RPKM_{TE}^{j}$ and $RPKM_{last_{CDS}}^{j}$ represent the normalized expression levels of the TE determined by IPAseek and the last coding exon of the gene, respectively, in each sample *j*. This methodology allows for comprehensive quantification of IPA isoform usage compared to the full-length isoform of the gene and facilitates direct comparison across multiple samples.

Benchmarking IPAseek Against Established IPA Detection Methods

To evaluate the performance of IPAseek in detecting IPA sites, we conducted a comparative analysis using four test samples with matched RNA-seq and 3' Seq data (**details provided in Supplementary_Table_S1**). IPA sites expressed at ≥5 TPM from the 3' Seq data served as the ground truth, while corresponding RNA-seq samples were processed using IPAseek to identify IPA sites.

To benchmark IPAseek against existing IPA detection methods, we analyzed the same test samples using InPACT and IPAFinder. Due to the resolution constraints of RNA-seq data, IPA sites detected within 350 bp of the ground truth were classified as true positives (TP). Ground truth IPA sites that were not detected in RNA-seq were considered false negatives (FN), while sites detected in RNA-seq but absent in the ground truth were categorized as false positives (FP). These false positives were further referred to as uncharacterized IPA sites in downstream analyses. This framework enabled a systematic comparison of IPAseek's accuracy and sensitivity relative to established methods, while accounting for bulk RNA-seq limitations.

To quantify performance, we calculated Precision, Recall, and F1-score for each method. The formulas used are as follows:

$$Precision = \frac{TP + FP}{TP}$$

$$Recall = \frac{TP + FN}{TP}$$

$$F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

This comprehensive evaluation allowed us to assess IPAseek's effectiveness in detecting IPA sites and compare its performance against InPACT and IPAFinder.

Annotation Assessment of Uncharacterized IPA Sites

All methods identified uncharacterized IPA sites in the test samples that were either absent or not expressed in the corresponding 3'-seq datasets. To investigate these uncharacterized sites further, we searched for potential 3'-end annotations in external databases, including PolyAsite, RefSeq, and Ensembl (O'Leary et al. 2016; Herrmann et al. 2020; Martin et al. 2023). As a first step, we examined the respective 3'-seq datasets to identify any potential annotations among the unexpressed candidates. For uncharacterized sites without annotation in the 3'-seq data, we searched within ±350 bp of the annotated 3' UTR end sites in PolyAsite, RefSeq, and Ensembl, following this order of priority. Uncharacterized sites that lacked annotation in any of these sources were categorized as unannotated.

IPA Detection and Quantification During Myeloid Progenitor to Megakaryocyte Differentiation

Bulk RNA-seq data from 16 samples, spanning eight time points (0 min, 30 min, 90 min, 180 min, 360 min, 1440 min, 2880 min, and 4320 min) with two replicates per time point, were obtained from GEO under accession number GSE213909. These samples capture the differentiation of K562 cells into megakaryocytes. The raw RNA-seq reads were aligned to the human reference genome (GRCh38) using STAR, and BAM files were filtered to retain only uniquely mapped reads. IPAseek was then applied to identify IPA sites, construct an atlas of IPA events, and calculate IPA site usage for each event across all 16 samples.

Experimental validation of IPA candidates

3'-RACE

The 3'-RACE was performed using the SMARTer® RACE 5'/3' Kit (Takara, 634858). Briefly, total cell RNA was extracted from K562 cells and converted to cDNA using SMARTer scribe reverse transcriptase. The specific genes were amplified from cDNA by gene specific primer in the exonic region of IPA along UPM followed by a PCR with nested gene specific primer. The amplified product was then cloned into pRACE vector using infusion cloning and sent to sequencing to obtain the 3'-end sequence of the RNA. A List of primer used is in the **Supplementary_Table_S3**. Further we generated a FASTA file containing the primer sequences and aligned them to the reference genome using Bowtie 2. The resulting BAM file was processed with BEDTools to create a BED file corresponding to primer positions. These coordinates were used to show the position of RACE primers on the Gviz tracks and have been provided in **Supplementary_Table_S3** along with the primer sequences.

Megakaryocyte differentiation from K562 cells

K562 cells were maintained in Iscove's Modified Dulbecco's Medium (IMDM, Gibco™, 12440053) supplemented with 10% Fetal Bovine serum (FBS, and Penicillin-Streptomycin (Penicillin-Streptomycin, Gibco™, 15140122). The differentiation of K562 to megakaryocyte was performed as per Bond et al (Bond et al. 2023). Briefly, 1×105 /ml of K562 cells were seeded in complete media and treated with 25nM of phorbol 12-myrisate 13-acetate (PMA) for six spanning time points (6, 12, 24, 48, 72 and 96 hrs). Media was changed after every 24 hours with supplementation of fresh PMA. Cells were harvested at 6,12, 24, 48, 72 and 96 hrs. Cells supplemented with 0.0025% DMSO for 96 hrs were used as control.

The differentiation of K562 to megakaryocyte was confirmed by ITGB3, a marker whose expression increases in megakaryocyte state. We used another gene *KLF1*, erythroid marker as a negative control whose expression decreases in megakaryocyte state. The expression of the genes was confirmed using qRT-PCR at 24, 48 and 72 hrs.

RNA Isolation and real time PCR

Total RNA was isolated using TRIzol reagent (MRC, TR118) by phenol-chloroform extraction method. Cell pellets were resuspended in TRIzol, phase-separated with chloroform and the RNA was precipitated with isopropanol followed by washing with 75% ethanol. The pellet was then air dried and resuspended in RNase-free water. The RNA was converted to cDNA using qScript™ cDNA SuperMix kit (Quantabio, 95048-100) as per manufacture instruction. PowerUp™ SYBR™ Green Master Mix (Applied Biosystems™, A25778) was used to qRT PCR. 25-50 ng RNA template was used for qRT PCR. Primers used for qRT PCR are listed in **Supplementary_Table_S3**. The relative expression was calculated using $2-\Delta\Delta Ct$ method and IPA site usage was calculated by using $2-\Delta\Delta Ct\_IPA\ /\ (2-\Delta\Delta Ct_{IPA}+\ 2-\Delta\Delta Ct_{FL})$ (Singh et al. 2018).

Construction and Analysis of an Immune Cell IPA Atlas

To construct an atlas of IPA events in immune cells, we downloaded bulk RNA-seq data from 42 samples under GEO accession number GSE184264 and 36 samples from GSE114922. These datasets encompass 10 immune cell-types, including naïve B cells, naïve CD4$^+$ T cells, naïve CD8$^+$ T cells, granulocytes, monocytes, natural killer cells, granulocytic precursor cells, monocytic precursor cells, erythroid precursor cells, and CD34$^+$ hematopoietic stem progenitor cells (**see Supplementary_Table_S4 for details).** Using IPAseek, we processed the 42 samples to identify IPA sites and construct the immune cell IPA atlas. Highly confident IPA events were defined as those with corresponding TEs expressed at RPKM ≥ 0.5 in and IPA site usage ≥ 10% in at least 10% samples of any sample group. Cell-type specific IPA events were identified by selecting highly confident IPA events detected in at least two samples associated with a given cell-type. We performed PCA followed by *k*-means clustering on the samples from GSE184264 and GSE114922 separately to identify the separation between the sample groups based on the IPA site usage.

For differential analysis of IPA site usage between two groups (e.g., cell-types or lineages), we performed a Wilcoxon rank-sum test for each IPA event and calculated the overall mean IPA site usage as well as group-specific means (group1_mean and group2_mean). $P-\text{values}$ were adjusted using the Benjamini-Hochberg FDR correction. IPA events with $P-\text{adj} < 0.05$ were classified as enriched. If group1_mean was higher, the event was labeled as "group1 enriched"; otherwise, it was labeled as "group2 enriched." For comparisons involving more than two groups (e.g., multiple cell-types), we applied the Kruskal-Wallis test for each IPA event, followed by FDR correction. Events with $P-\text{adj} < 0.05$ were identified as enriched IPA events.

Relation between IPA and DNA methylation in immune cells

Identification of Matched RNA-seq and BS-seq Samples

To explore the relationship between IPA site selection and DNA methylation, we curated 198 bulk RNA-seq and 189 BS-seq samples (RRBS/WGBS) from ENCODE (primary cells), dbGaP (phs001027), and GEO (GSE165305, GSE214980, GSE184314, GSE156563, GSE130582, GSE66117, GSE173790, GSE128269, GSE193201). These represented 12 immune cell-types: B cells, NK cells, T cells, CD14$^+$ cells, PTCL, CD4$^+$ cells, B-ALL, T helper cells, HSPCs, PBMC-reprogrammed iMSCs, AML, and NBM (**see Supplementary_Table_S6**).

IPA Sites Atlas Construction

IPAseek identified IPA events through intron preprocessing, STAR alignment with RPKM normalization, PELT changepoint detection, and validation using splice read exclusion and RPKM thresholds. Events within 100 bp were merged, and confidence was assigned based on detection frequency.

Methylation Data Processing Pipeline

BS-seq reads were trimmed (trim-galore v0.6.10), aligned to bisulfite-converted GRCh38 (Bismark v0.24.1), deduplicated, and methylation calls extracted (Krueger and Andrews 2011).

Methylation Atlas Generation

Using edgeR v4.4.2 (Robinson et al. 2010; Chen et al. 2018), sites were filtered (≥10 reads, ≥50% methylation in ≥2 samples) and annotated to GRCh38 gene bodies. Methylation levels were calculated as proportion of methylated reads. Next we the PAS sites into used and unused and performed the Integrative Analysis of Methylation Patterns around IPA Sites using Rolling Window and Equal-Tile Analysis (Refer to Supplementary Methods for details)

Data sets

A detailed description of datasets used in study is provided in **Supplementary_Table_S1, S4, S5&S6**. 3'-seq and bulk RNA-seq datasets for testing and benchmarking IPAseek were downloaded from GEO accession number GSE111793 & GSE111310 (Test Sample 2-18, Test Sample 1&2 provided as raw data on IPAseek GitHub repository). Bulk RNA-seq datasets for studying the temporal dynamics of IPA during megakaryocyte differentiation was downloaded from GEO accession number GSE213909. Bulk RNA-seq datasets from GSE184264 and GSE114922 were used for making IPA atlas for immune cells. For understanding the interplay between IPA site usage and DNA methylation, the RNA-seq and matched BS-seq (RRBS or WGBS) samples were used from ENCODE, GEO (GSE165305, GSE214980, GSE184314, GSE156563, GSE130582, GSE66117, GSE173790, GSE128269, GSE193201), dbGaP (phs001027).

Software availability

IPAseek is an open-source method available on GitHub repository (**https://github.com/isinghlab/IPAseek.git**) and as Supplemental Code.

Competing interest statement

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by 1R21NS121945, 1R01CA282251, CPRIT RP230204, and Texas A&M Health Science Center Seedling grant awarded to IS. We thank all Singh Lab members in the group for their assistance and constructive suggestions. We also gratefully acknowledge the technical support by the High-Performance Research Computing resources at Texas A&M

University. Artificial intelligence-based tools were used exclusively for language editing and text refinement.

Author contributions

Authors and Affiliations

Department of Cell Biology & Genetics, Texas A&M University Health Science Center, Bryan, TX

Richa Rashmi, Pranita Borkar, Sumana Mallick, Taylor Hubbs, Ari Aviles, Daniel Chung and Irtisha Singh

Department of Biomedical Engineering, Texas A&M University, College Station, TX

Abhinaya Muruganandham, Sumana Mallick, Irtisha Singh

Interdisciplinary Program in Genetics and Genomics, Texas A&M University, College Station, TX

Ari Aviles, Irtisha Singh

Contributions

IS and RR conceived and designed the study. IS supervised the study and data analysis. RR implemented the idea, wrote the original code, and analyzed the data. AM helped in data analysis. PB and SM performed the 3'-RACE experiments and performed the K562 differentiation experiments respectively and analyzed the outcoming data. AA and DC performed the RNA isolation and cDNA conversion for the K562 differentiation study. TH performed the qRT-PCR for the K562 differentiation study. RR and IS wrote the manuscript with input from all authors. All authors approved the final version submitted.

Corresponding author

Correspondence to Irtisha Singh.

References

Uncategorized References

Arefeen A, Liu J, Xiao X, Jiang T. 2018. TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* **34**: 2521–2529. doi:10.1093/bioinformatics/bty110.

Berkovits BD, Mayr C. 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**: 363–367. doi:10.1038/nature14321.

Bond ML, Davis ES, Quiroga IY, Dey A, Kiran M, Love MI, Won H, Phanstiel DH. 2023. Chromatin loop dynamics during cellular differentiation are associated with changes to both anchor and internal regulatory features. *Genome Res* **33**: 1258–1268. doi:10.1101/gr.277397.122.

Chen Y, Pal B, Visvader JE, Smyth GK. 2018. Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR. *F1000Research* **6** doi:10.12688/f1000research.13196.2.

Cheng LC, Zheng D, Baljinnyam E, Sun F, Ogami K, Yeung PL, Hoque M, Lu C-W, Manley JL, Tian B et al. 2020. Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation. *Nature Communications 2020 11:1* **11** doi:10.1038/s41467-020-16959-2.

Cheng X, Jiang G, Zhou X, Wang J, Zhao Z, Zhang J, Ni T. 2024. The landscape and clinical relevance of intronic polyadenylation in human cancers. *J Genet Genomics* **51**: 1030–1039. doi:10.1016/j.jgg.2024.04.014.

Cowley M, Wood AJ, Bohm S, Schulz R, Oakey RJ. 2012. Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. *Nucleic Acids Res* **40**: 8917–8926. doi:10.1093/nar/gks654.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173–1183. doi:10.1101/gr.132563.111.

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

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**: 313–319. doi:10.1016/0092-8674(80)90617-0.

Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14**: 496–506. doi:10.1038/nrg3482.

Fink EE, Zhang Y, Santo B, Siddavatam A, Ou R, Nanavaty V, Lee BH, Ting AH. 2025. Heat shock induces alternative polyadenylation through dynamic DNA methylation and chromatin looping. *Cell Stress Chaperones* **30**: 100084. doi:10.1016/j.cstres.2025.100084.

Gabel AM, Belleville AE, Thomas JD, McKellar SA, Nicholas TR, Banjo T, Crosse EI, Bradley RK. 2024. Multiplexed screening reveals how cancer-specific alternative polyadenylation shapes tumor growth in vivo. *Nat Commun* **15**: 959. doi:10.1038/s41467-024-44931-x.

Griffiths-Jones S. 2004. The microRNA Registry. *Nucleic Acids Res* **32**: D109–111. doi:10.1093/nar/gkh023.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**: D140–144. doi:10.1093/nar/gkj112.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**: D154–158. doi:10.1093/nar/gkm952.

Gruber AJ, Schmidt R, Ghosh S, Martin G, Gruber AR, van Nimwegen E, Zavolan M. 2018. Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biol* **19**: 44. doi:10.1186/s13059-018-1415-3.

Ha KCH, Blencowe BJ, Morris Q. 2018. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol* **19**: 45. doi:10.1186/s13059-018-1414-4.

Haynes K, Eckley IA, Fearnhead P. 2017. Computationally Efficient Changepoint Detection for a Range of Penalties. *Journal of Computational and Graphical Statistics* **26**: 134–143. doi:10.1080/10618600.2015.1116445.

Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2020. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**: D174–D179. doi:10.1093/nar/gkz918.

Hunt MC, Siponen MI, Alexson SE. 2012. The emerging role of acyl-CoA thioesterases and acyltransferases in regulating peroxisomal lipid metabolism. *Biochim Biophys Acta* **1822**: 1397–1410. doi:10.1016/j.bbadis.2012.03.009.

Jia J, Fan H, Wan X, Fang Y, Li Z, Tang Y, Zhang Y, Huang J, Fang D. 2024. FUS reads histone H3K36me3 to regulate alternative polyadenylation. *Nucleic Acids Res* **52**: 5549–5571. doi:10.1093/nar/gkae184.

Killick R, Fearnhead P, Eckley IA. 2012. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association* **107**: 1590–1598. doi:10.1080/01621459.2012.737745.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572. doi:10.1093/bioinformatics/btr167.

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

Li Y, Schaefke B, Zou X, Zhang M, Heyd F, Sun W, Zhang B, Li G, Liang W, He Y et al. 2020. Pan-tissue analysis of allelic alternative polyadenylation suggests widespread functional regulation. *Mol Syst Biol* **16**: e9367. doi:10.15252/msb.20199367.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**: 2380–2396. doi:10.1101/gad.229328.113.

Liu X, Chen H, Li Z, Yang X, Jin W, Wang Y, Zheng J, Li L, Xuan C, Yuan J et al. 2024. InPACT: a computational method for accurate characterization of intronic polyadenylation from RNA sequencing data. *Nat Commun* **15**: 2583. doi:10.1038/s41467-024-46875-8.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. doi:10.1186/s13059-014-0550-8.

Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J et al. 2023. Ensembl 2023. *Nucleic Acids Res* **51**: D933–D941. doi:10.1093/nar/gkac958.

Mayr C. 2019. What Are 3' UTRs Doing? Cold Spring Harb Perspect Biol 11 doi:10.1101/cshperspect.a034728.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673–684. doi:10.1016/j.cell.2009.06.016.

Middleton R, Gao D, Thomas A, Singh B, Au A, Wong JJ, Bomane A, Cosson B, Eyras E, Rasko JE et al. 2017. IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol 18: 51. doi:10.1186/s13059-017-1184-4.

Monteuuis G, Wong JJL, Bailey CG, Schmitz U, Rasko JEJ. 2019. The changing paradigm of intron retention: regulation, ramifications and recipes. Nucleic Acids Res 47: 11497–11513. doi:10.1093/nar/gkz1068.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628. doi:10.1038/nmeth.1226.

Nanavaty V, Abrash EW, Hong C, Park S, Fink EE, Li Z, Sweet TJ, Bhasin JM, Singuri S, Lee BH et al. 2020. DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex. *Mol Cell* **78**: 752–764 e756. doi:10.1016/j.molcel.2020.03.024.

Neve J, Patel R, Wang Z, Louey A, Furger AM. 2017. Cleavage and polyadenylation: Ending the message expands gene regulation. RNA Biol 14: 865–890. doi:10.1080/15476286.2017.1306171.

Ni TK, Kuperwasser C. 2016. Premature polyadenylation of MAGI3 produces a dominantly-acting oncogene in human breast cancer. Elife 5 doi:10.7554/eLife.14730.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–745. doi:10.1093/nar/gkv1189.

Plessner M, Thiele L, Hofhuis J, Thoms S. 2024. Tissue-specific roles of peroxisomes revealed by expression meta-analysis. *Biol Direct* **19**: 14. doi:10.1186/s13062-024-00458-1.

Proudfoot NJ. 2011. Ending the message: poly(A) signals then and now. Genes Dev 25: 1770–1782. doi:10.1101/gad.17268411.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi:10.1093/bioinformatics/btp616.

Samur MK, Minvielle S, Gulla A, Fulciniti M, Cleynen A, Aktas Samur A, Szalat R, Shammas M, Magrangeas F, Tai YT et al. 2018. Long intergenic non-coding RNAs have an independent impact on survival in multiple myeloma. Leukemia 32: 2626–2635. doi:10.1038/s41375-018-0116-y.

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

Scotto-Lavino E, Du G, Frohman MA. 2006. 3' end cDNA amplification using classic RACE. *Nat Protoc* **1**: 2742–2745. doi:10.1038/nprot.2006.481.

Shenker S, Miura P, Sanfilippo P, Lai EC. 2015. IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA* **21**: 14–27. doi:10.1261/rna.046037.114.

Singh I, Lee SH, Sperling AS, Samur MK, Tai YT, Fulciniti M, Munshi NC, Mayr C, Leslie CS. 2018. Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat Commun* **9**: 1716. doi:10.1038/s41467-018-04112-z.

Smith LM. 2019. Epigenetic Regulation of mRNA Polyadenylation Site Selection. *Plant Physiol* **180**: 7–9. doi:10.1104/pp.19.00374.

Sun J, Kim J-Y, Jun S, Park M, de Jong E, Chang J-W, Cheng S, Fan D, Chen Y, Griffin TJ et al. 2024. Dichotomous intronic polyadenylation profiles reveal multifaceted gene functions in the pan-cancer transcriptome. *Experimental & Molecular Medicine 2024 56:10* **56** doi:10.1038/s12276-024-01289-w.

Takagaki Y, Manley JL. 1998. Levels of Polyadenylation Factor CstF-64 Control IgM Heavy Chain mRNA Accumulation and Other Events Associated with B Cell Differentiation. *Molecular Cell* **2** doi:10.1016/S1097-2765(00)80291-9.

Thomas PE, Wu X, Liu M, Gaffney B, Ji G, Li QQ, Hunt AG. 2012. Genome-Wide Control of Polyadenylation Site Choice by CPSF30 in Arabidopsis. *The Plant Cell* **24** doi:10.1105/tpc.112.096107.

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

Tian B, Pan Z, Lee JY. 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* **17**: 156–165. doi:10.1101/gr.5532707.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476. doi:10.1038/nature07509.

Weber MJ. 2005. New human and mouse microRNA genes found by homology search. *FEBS J* **272**: 59–73. doi:10.1111/j.1432-1033.2004.04389.x.

Wood AJ, Schulz R, Woodfine K, Koltowska K, Beechey CV, Peters J, Bourc'his D, Oakey RJ. 2008. Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev* **22**: 1141–1146. doi:10.1101/gad.473408.

Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**: 5274. doi:10.1038/ncomms6274.

Zhao Z, Xu Q, Wei R, Wang W, Ding D, Yang Y, Yao J, Zhang L, Hu YQ, Wei G et al. 2021. Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res* **31**: 2095–2106. doi:10.1101/gr.271627.120.

Fig. 1: Changepoint Detection of IPA Sites by PELT.

A. Top - Bulk RNA-seq coverage (raw read count of uniquely mapping reads) over a genomic locus annotated as intron by RefSeq. The sashimi plot shows the splicing complexity at the locus. Bottom - Peaks (raw read count of mapped reads at the locus) detected by 3'-seq, a high throughput 3'-end detection and quantification approach. This is a representation of a composite IPA event. The gene locus presented here is located on the sense strand of DNA.

B. Elbow plot demonstrates the number of changepoints (x-axis) detected by PELT over a range of penalties (y-axis) over the locus shown in A. This plot shows detection of one change point generated due to 3'-end formation.

C. Plot shows the read coverage over the intron in A. It highlights the two segments detected by PELT with the maximum difference in mean of the segments.

D. As A but shows the locus of a skipped IPA event. The gene locus presented here is located on the sense strand of DNA.

E. As B but shows detection of two change points for the locus shown in D. One change point explains the splicing and the second change point highlights the 3'-end formation.

F. As in C but for the locus shown in D. PELT detects three segments with different means using the read coverage over the intron shown in D.

Fig. 2: Comprehensive analysis of IPA events detected by IPAseek.

A. Same as 1A but a composite IPA event only detected by IPAseek. The highlighted region shows the new terminal exon (TE), where the termination site was detected as a change point by IPAseek. The gene locus presented here is located on the sense strand of DNA.

B. Similar visualization as in A but depicting a skipped IPA event detected only by IPAseek. Top track here indicates the position of the two nested primers used for 3'-RACE validation. The blue highlighted region shows the TE detected by IPAseek and the pink highlighted region shows the cleavage site detected by 3'-RACE (±20 nt). The gene locus presented here is located on the sense strand of DNA.

C. Validation of 3'-end detected by IPAseek in B using 3'-RACE. Total RNA extracted from K562 cells were used to perform 3'-RACE.

Fig. 3: Comparative Analysis of IPAseek Detected IPA Events with Existing Methods.

A. A comprehensive evaluation of the performance metrics for IPAseek, InPACT, and IPAFinder in detecting IPA events across multiple 6 test samples (ordered by sample quality). The analysis employs three key metrics: Precision, Recall, and F1-Score. The IPA 3'-ends detected by 3'-seq (expressed at ≥ 5 TPM) were utilized as ground truth. Precision measures the proportion of correctly identified IPA events among all detected events, while recall quantifies the fraction of true IPA events successfully identified by each method. The F1-Score, the harmonic mean of precision and recall, provides a balanced measure of overall performance.

B. Comparison of bulk RNA-seq expression levels of TEs, represented as $\log_2$ (RPKM + 1), between IPA events detected by both IPAseek and 3'-seq (True Positives) and those identified exclusively by 3'-seq (False Negatives). Statistical significance of the difference in TE expression between True Positives and False Negatives was assessed using the Kolmogorov-Smirnov test ($P-\mathrm{value} < 0.001$ of Test sample 1,2,4,8 and $P-\mathrm{value} < 0.01$ of Test sample 14).

C. This figure presents a quantitative comparison of previously uncharacterized IPA events (False Positives) identified by IPAseek, InPACT, and IPAFinder across the 6 test Samples (ordered by sample quality). Previously uncharacterized IPAs are defined as those detected using RNA-seq coverage profiles that either are not present in the 3'-seq atlas or have low expression (≤ 5 TPM) in the 3'-seq atlas. The color-coded segments within each bar represent the distribution of annotation sources for these uncharacterized IPA sites, including the 3'-seq atlas (expressed ≤ 5 TPM), PolyASite database, RefSeq, and Ensembl annotations, prioritized in that order. The grey region indicates IPA events detected by each method that lack any prior annotation.

D. Proportion of uncharacterized IPA events without annotations detected in 6 Test Samples (ordered by sample quality, grey events in C for IPAseek) with a canonical/non-canonical PAS within $\pm$150nt of the detected IPA site.

E. Density plot of the distance of the nearest PAS from the IPA cleavage site detected by IPAseek.

F. Same as 1A. Shows an uncharacterized composite IPA isoform without annotation detected by IPAseek but not by InPACT and IPAFinder. The gene locus presented here is located on the sense strand of DNA.

G. Same as 1A. Shows an uncharacterized skipped IPA isoform without annotation detected by IPAseek but not by InPACT and IPAFinder. The gene locus presented here is located on the antisense strand of DNA.

Fig. 4: Temporal Analysis of IPA During Myeloid Progenitor to Megakaryocyte Differentiation

A. IPA sites with differential usage (≥ 10% usage difference) with respect to full-length mRNA in myeloid progenitor cells (K562) to megakaryocytic differentiation (n = 166). The sites were determined by the calculating the difference in IPA site usage of all the IPA events, using the two early and two end differentiation time points. Each row is a unique IPA isoform, while columns are samples captured at different time points of the differentiation process. The color scheme indicates regulation status, with higher usage shown in orange and lower usage in blue. IPA site usage is quantified on a 0-1 scale relative to full-length isoform usage, with color intensity reflecting the degree of IPA site usage.

B. RNA-seq read coverage (in TPM) over *ACOT8* locus across different timepoints of myeloid progenitor (K562) to megakaryocytic differentiation. The read coverage was determined using uniquely mapping reads. Topmost track indicates the position of the two nested primers used for 3'-RACE validation. The blue highlighted region shows the TE determined by IPAseek while pink highlighted region shows the cleavage site detected by 3'-RACE (±20 nt). The gene locus presented here is located on the antisense strand of DNA.

C. Like B but showing *PEX13* locus. The gene locus presented here is located on the sense strand of DNA.

D. Validation of 3'-end detected by IPAseek in B & C using 3'-RACE. Total cell RNA extracted from K562 cells were used to perform 3'-RACE.

E. Morphological assessment of myeloid progenitor cells. Phase contrast microscopy (20× magnification; BioTek Lionheart FX Automated Microscope) captured progressive morphological changes during PMA-induced megakaryocytic differentiation of myeloid progenitor cells. Columns represent treatment groups: untreated control (left), 0.0025% DMSO vehicle (middle), and 25 nM PMA-treated cells (right). Rows correspond to time points: 24 h (top), 48 h (middle), and 72 h (bottom). PMA-treated cells exhibited hallmark differentiation features including increased cell size, cytoplasmic expansion, and enhanced substrate adhesion compared to controls. Scale bars: 10 μm.

F. IPA isoform usage in *ACOT8* during myeloid progenitor differentiation to megakaryocytes. The left panel shows IPA site usage measured across eight time points (0 min to 4320 min) following PMA treatment, quantified using IPAseek analysis of bulk RNA-seq data from GSE213909. The right panel presents IPA isoform usage, identified by 3'-RACE, measured at six time points (6 h to 96 h) post-PMA treatment or vehicle control (DMSO 96 h), validated by qRT-PCR. Together, these plots demonstrate an increase in *ACOT8* IPA isoform usage throughout differentiation.

G. Same as F but for *PEX13*. It demonstrates a decrease in *PEX13* IPA isoform usage throughout differentiation.

Fig. 5: Immune cell IPA atlas and functional enrichment analysis.

A. Principal Component Analysis (PCA) of differentiated immune cell-types from the GEO dataset GSE184264 based on IPA site usage. Three clusters identified through *k*-means clustering are indicated by dotted enclosures. The plot reveals clear segregation between myeloid (Clusters 1 and 2) and lymphoid (Clusters 2 and 3) lineages, reflecting distinct lineage-specific IPA patterns. The analysis includes 4,994 IPA events across 39 samples.

B. PCA of undifferentiated immune cell-types from GEO dataset GSE114922 based on IPA site usage. Four clusters identified through *k*-means clustering are indicated by dotted enclosures. The analysis reveals separation between individual cell-types, emphasizing their unique IPA signatures (No. of IPA = 592, No. of Samples = 108).

C. Heatmap of IPA site usage across differentiated immune cell-types, showcasing myeloid and lymphoid specific enrichment patterns. Each row represents a unique IPA isoform (n = 1,863), while columns correspond to individual immune cell samples (n = 42). Rows are color-coded to indicate enrichment status: myeloid enriched IPAs (seafoam green) and lymphoid enriched IPAs (dark orange). IPA site usage is quantified on a 0-1 scale relative to full-length isoform usage, with color intensity reflecting the degree of IPA site usage. Statistical significance was determined using the Wilcoxon rank-sum test with Benjamini-Hochberg FDR correction ($P - \text{adj} < 0.05$).

D. Heatmap of IPA site usage across undifferentiated immune cell-types, focusing on the cell-types shown in (B). Each row represents a unique IPA isoform (n = 431), and columns denote individual immune cell samples (n = 108). Statistical significance for enrichment in specific cell-types was assessed using the Kruskal-Wallis test with Benjamini-Hochberg FDR correction ($P - \text{adj} < 0.05$). IPA site usage is displayed on a 0-1 scale, with color intensity indicating the degree of IPA site usage.

E. Differential IPA site usage between myeloid and lymphoid lineages, with *RPN1* shown as an example of a lineage-specific IPA event. Bulk RNA-seq read coverage (in TPM) over the *RPN1* locus is visualized across different immune cell-types. Read coverage was calculated using uniquely mapping reads.

Fig. 6: Epigenetic Regulation of IPA in Immune Cells.

A. Integrated Methylation Analysis at IPA Sites. Metagene analysis of DNA methylation patterns within ±2500 bp of PAS revealed distinct epigenetic landscapes between used (orange) and unused (navy) PAS loci. Used PAS sites exhibited significantly higher methylation levels across all samples compared to unused PAS (Kolmogorov-Smirnov test, $P - \text{value} \leq 9.88 \times 10^{-270}$), with mean methylation differences persisting across the entire genomic window. This analysis encompassed 910 used PAS sites, 1,820 unused PAS sites, and ~5.7 million genomic methylation sites.

B. Cell-type specific Methylation State around IPA Sites in B cells. Metagene analysis of DNA methylation within ±2500 bp of PAS in B cells revealed distinct epigenetic profiles between used (orange) and unused (navy) PAS loci. Used PAS sites exhibited higher methylation levels compared to unused PAS sites across the entire genomic window (Kolmogorov-

Smirnov test, $P - \text{value} \leq 5.14 \times 10^{-177}$).This analysis included 679 used PAS sites, 1,358 unused PAS sites, and ~3.4 million.

C. Cell-Type Specific Methylation Patterns at IPA Sites in CD14⁺ Cells. Metagene analysis of DNA methylation within ±2500 bp of PAS in CD14⁺ cells revealed that used PAS sites (orange) are consistently flanked by higher methylation levels compared to unused PAS sites (navy) (Kolmogorov-Smirnov test, $P - \text{value} \leq 3.94 \times 10^{-244}$). This analysis included 325 used PAS sites, 650 unused PAS sites, and ~2.95M methylation sites.

D. DNA Methylation Landscapes at IPA Sites in AML. Metagene analysis of DNA methylation within ±2500 bp of PAS sites in AML demonstrates that used PAS sites (orange) are consistently associated with higher levels of DNA methylation compared to unused PAS sites (navy) (Kolmogorov-Smirnov test, P-value < 10^{-244}). This analysis encompassed 552 used PAS sites, 1,104 unused PAS sites, and ~2.4 million methylation sites.

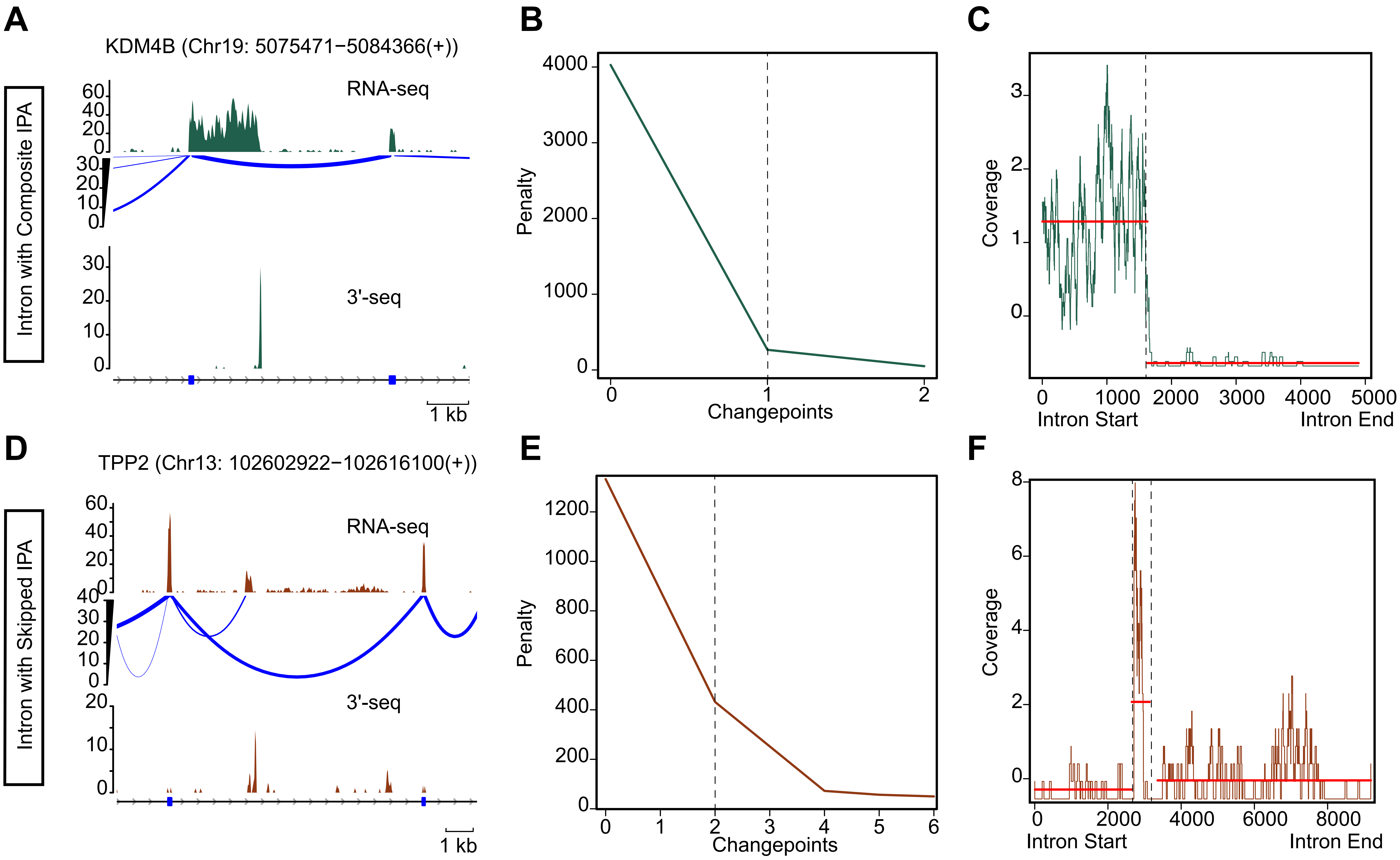
E. Rolling Window Analysis of DNA methylation in IPA and Non-IPA introns. Mean DNA methylation levels (y-axis) were calculated for IPA introns (violet), non-IPA introns within IPA genes (yellow), and introns from genes without IPA events (brown) using a sliding window of 100 bp along the intron length (x-axis) with a 50 bp step size. This analysis included 696 IPA introns, 1,758 non-IPA introns from IPA genes, and 1,303 introns from non-IPA genes, encompassing ~5.7 million methylation sites in total. IPA introns consistently showed significantly higher methylation levels across their entire length compared to both non-IPA introns and introns from non-IPA genes (Kolmogorov-Smirnov test, $P - \text{value} \leq 2 \times 10^{-16}$ for both comparisons).

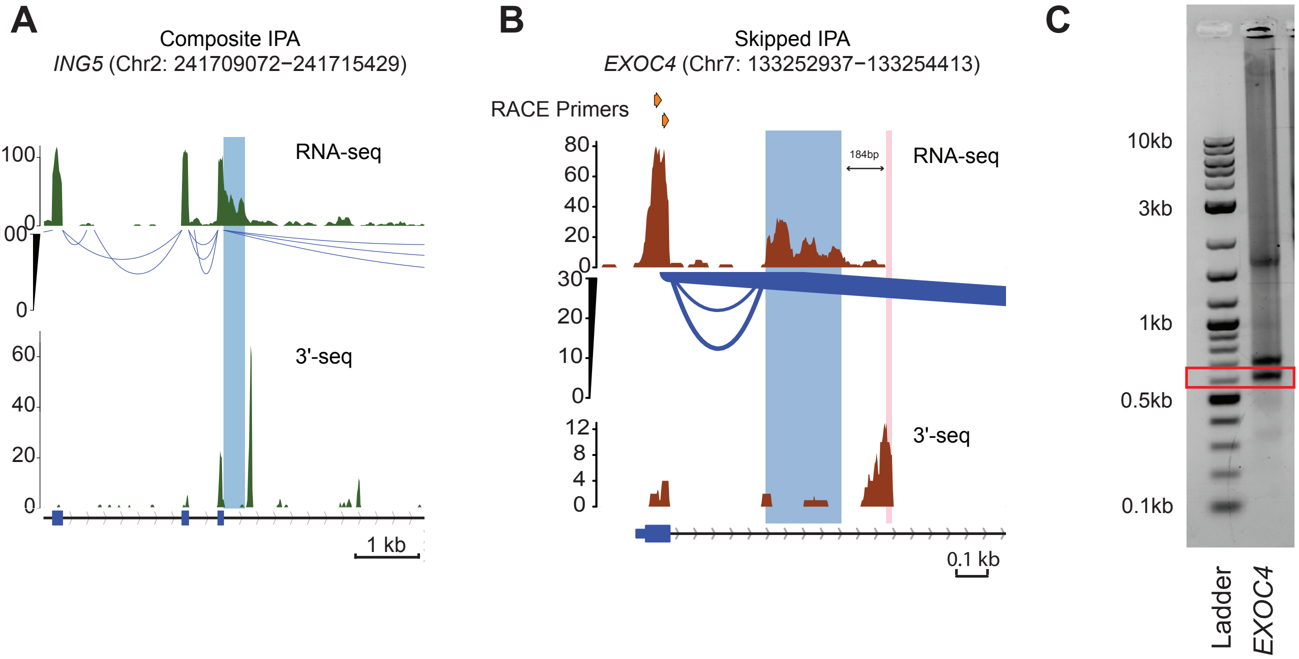
F. Heatmap of IPA site usage and mean methylation in the introns with differentially methylated sites (GSE184314). The left heatmap shows the mean methylation in the introns coming from differentially methylated sites and right heatmap shows the IPA site usage in the corresponding introns. Each row represents a unique intron (n = 101), while columns correspond to individual patient samples (n = 10). Rows are color-coded to indicate the differential methylation status in the introns: Hypermethylation (Green) (n = 38) and Hypomethylation (Orange) (n = 63). Columns are color-coded to indicate the sample groups: Control (Blue) (n = 5) and Diseased (Pink) (n = 5). Methylation is quantified on a 0-1 scale (blue) with color intensity reflecting the degree of methylation and IPA site usage is quantified on a 0-1 scale (red) relative to full-length isoform usage, with color intensity reflecting the degree of IPA site usage.

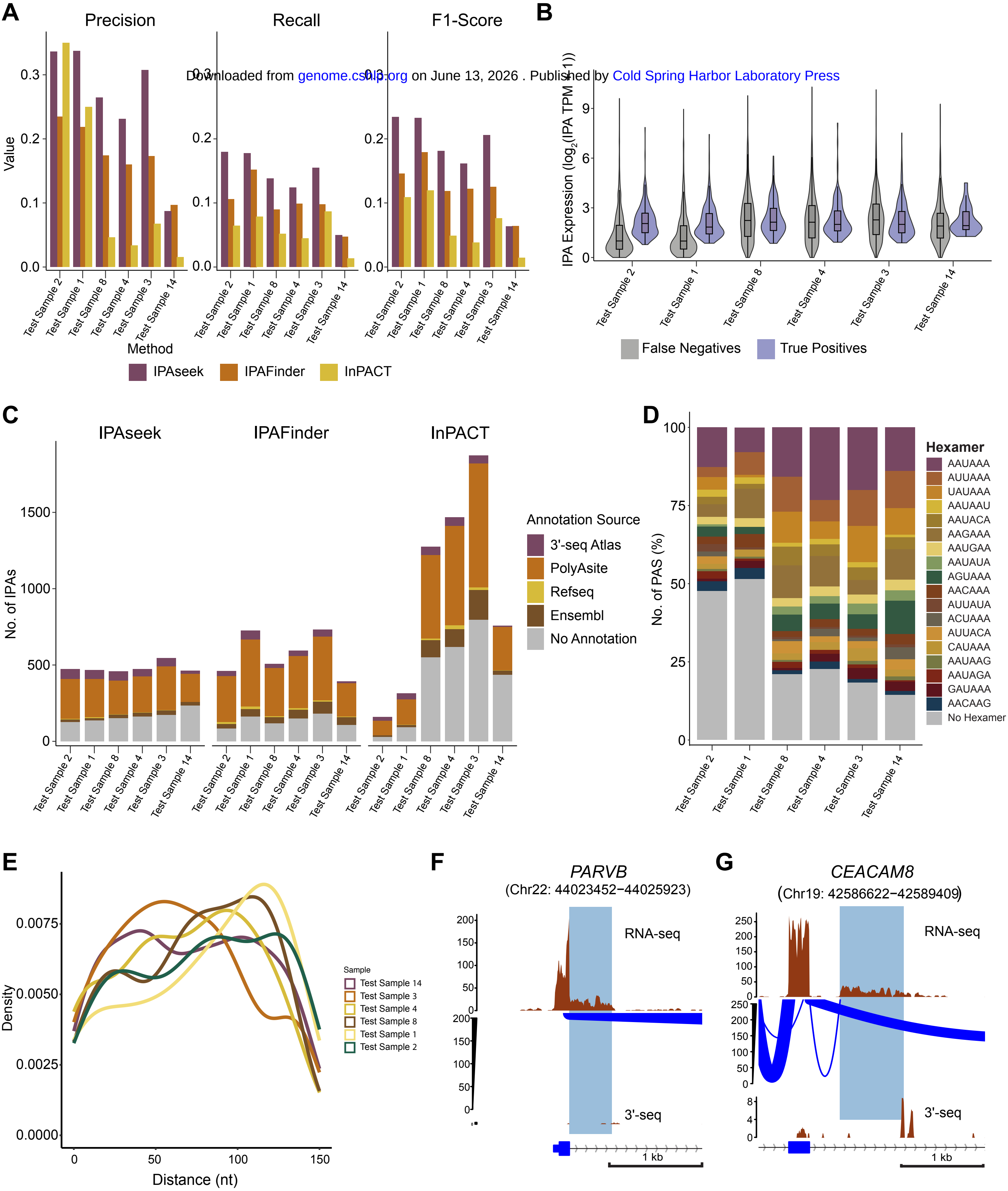
G. Density Plot of IPA site usage in Introns with Differentially Methylated Sites. Density plots compare IPA site usage in introns harboring differentially methylated sites, separated into hypermethylated (left panel) and hypomethylated (right panel) groups as identified in **Fig. 6F**. In the hypermethylated group (n = 380), IPA site usage differs significantly between Control (blue) and Diseased (pink) samples (Wilcoxon paired test, $P - \text{value} \leq 3 \times 10^{-10}$), whereas in the hypomethylated group (n = 630), no significant difference is observed (Wilcoxon paired test, $P - \text{value} < 0.56$).

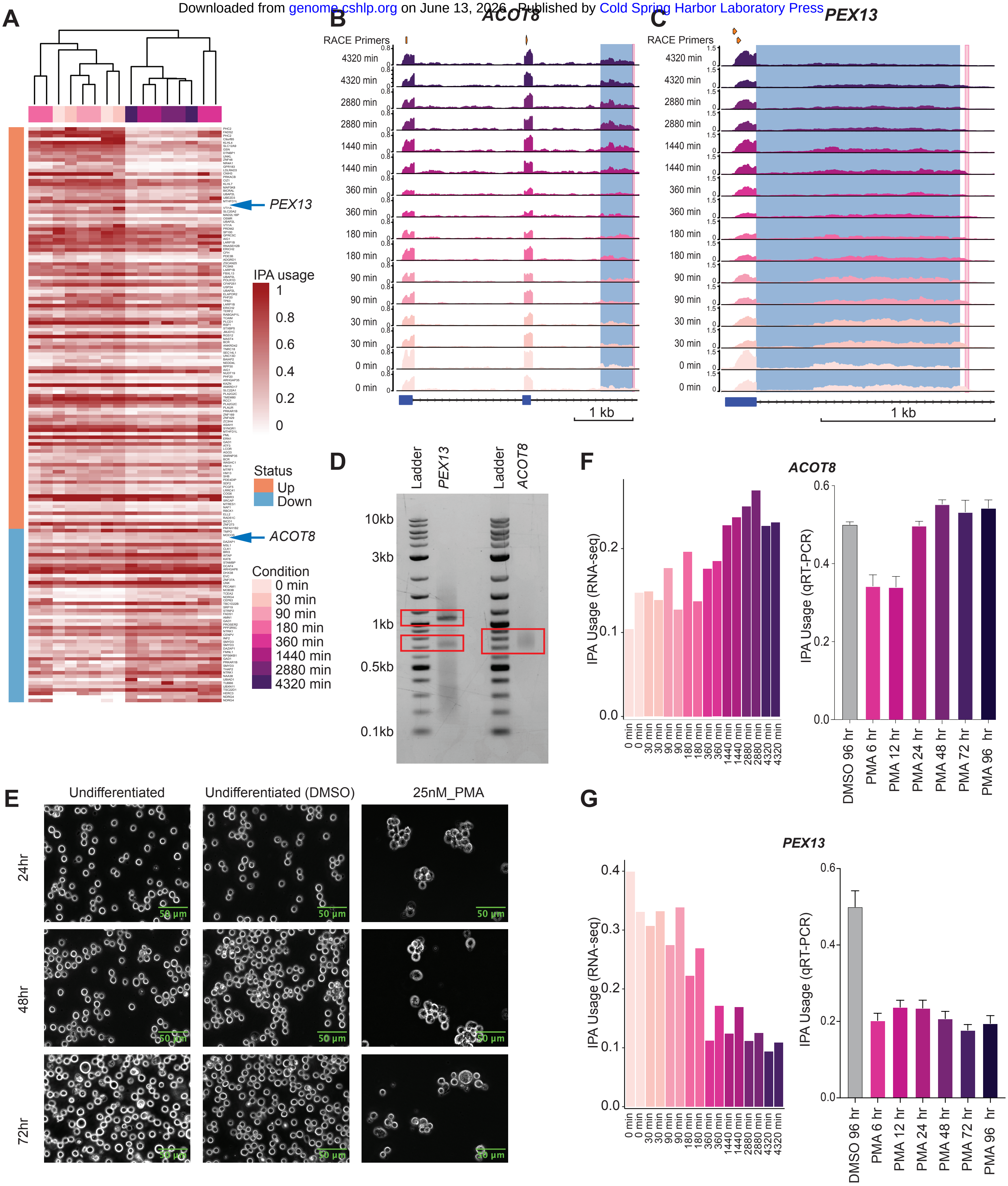
H. Heatmap of IPA site usage and mean methylation in the introns with differentially used IPA sites (GSE184314). The left heatmap shows the IPA site usage in the introns coming from differentially used IPA sites. and right heatmap shows the methylation in the corresponding introns. Each row represents a unique intron (n = 102), while columns correspond to individual patient samples (n = 10). Rows are color-coded to indicate the differential IPA site usage status in the introns: highly used (dark green; n = 36) and lowly used (brown; n = 66). Columns are color-coded to indicate the sample groups: Control (blue; n = 5) and Diseased (pink; n = 5). Methylation is quantified on a 0-1 scale (blue) with color intensity reflecting the degree of methylation and IPA site usage is quantified on a 0-1 scale (red) relative to full-length isoform usage, with color intensity reflecting the degree of IPA site usage.

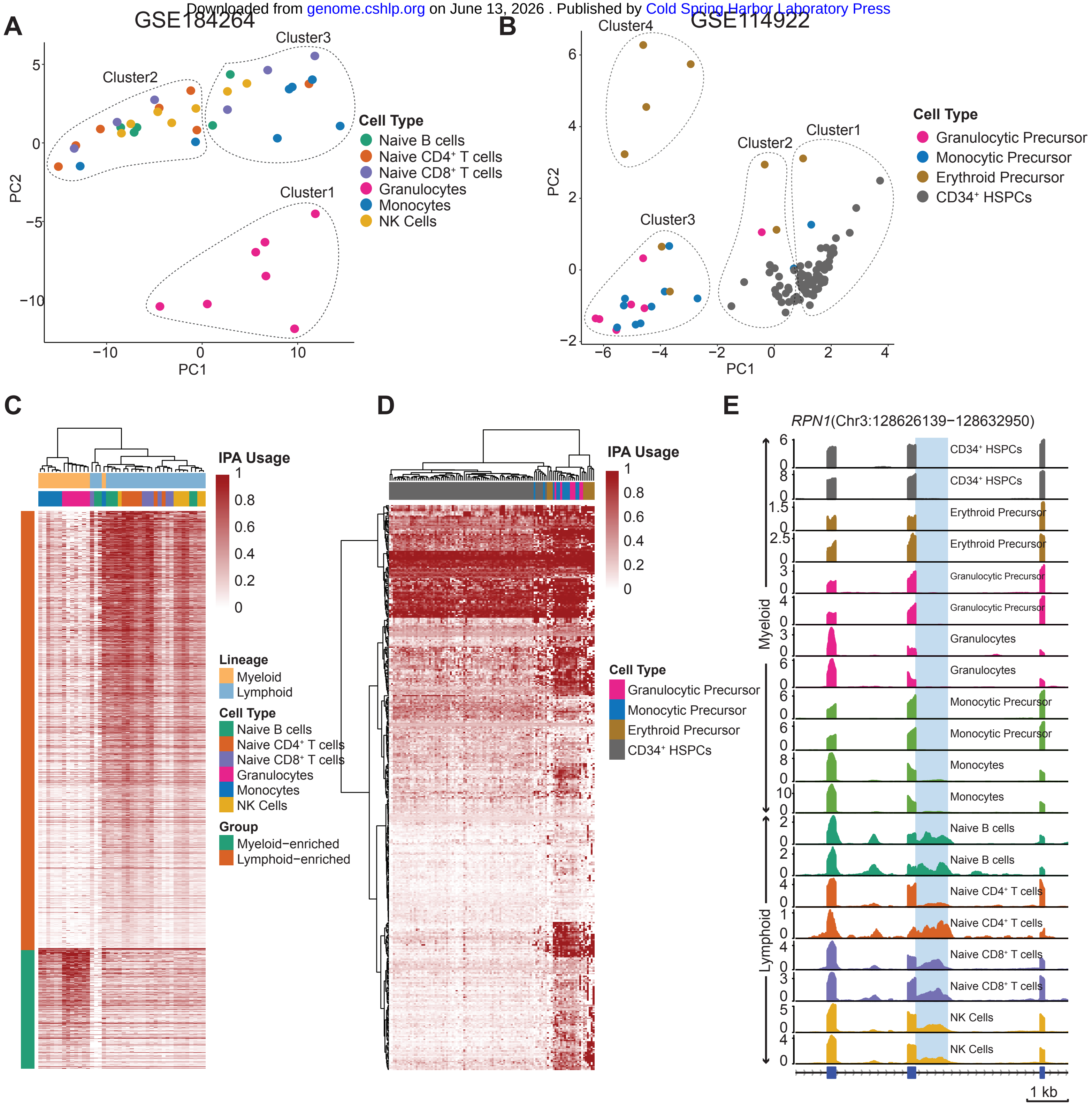
I. Density Plot of Methylation Levels in Introns with Differentially Used IPA Sites. Density plots illustrate methylation levels in introns containing differentially used IPA sites, divided into lowly used (left panel) and highly used (right panel) groups as defined in **Fig. 6H**. For lowly used IPA sites (n = 660), a modest but significant difference in methylation is observed between control (blue) and diseased (pink) samples (Wilcoxon paired test, $P - \text{value} < 0.15$). In contrast, highly used IPA sites (n = 360) show a pronounced and highly significant methylation difference between groups (Wilcoxon paired test, $P - \text{value} \leq 2.1 \times 10^{-10}$).

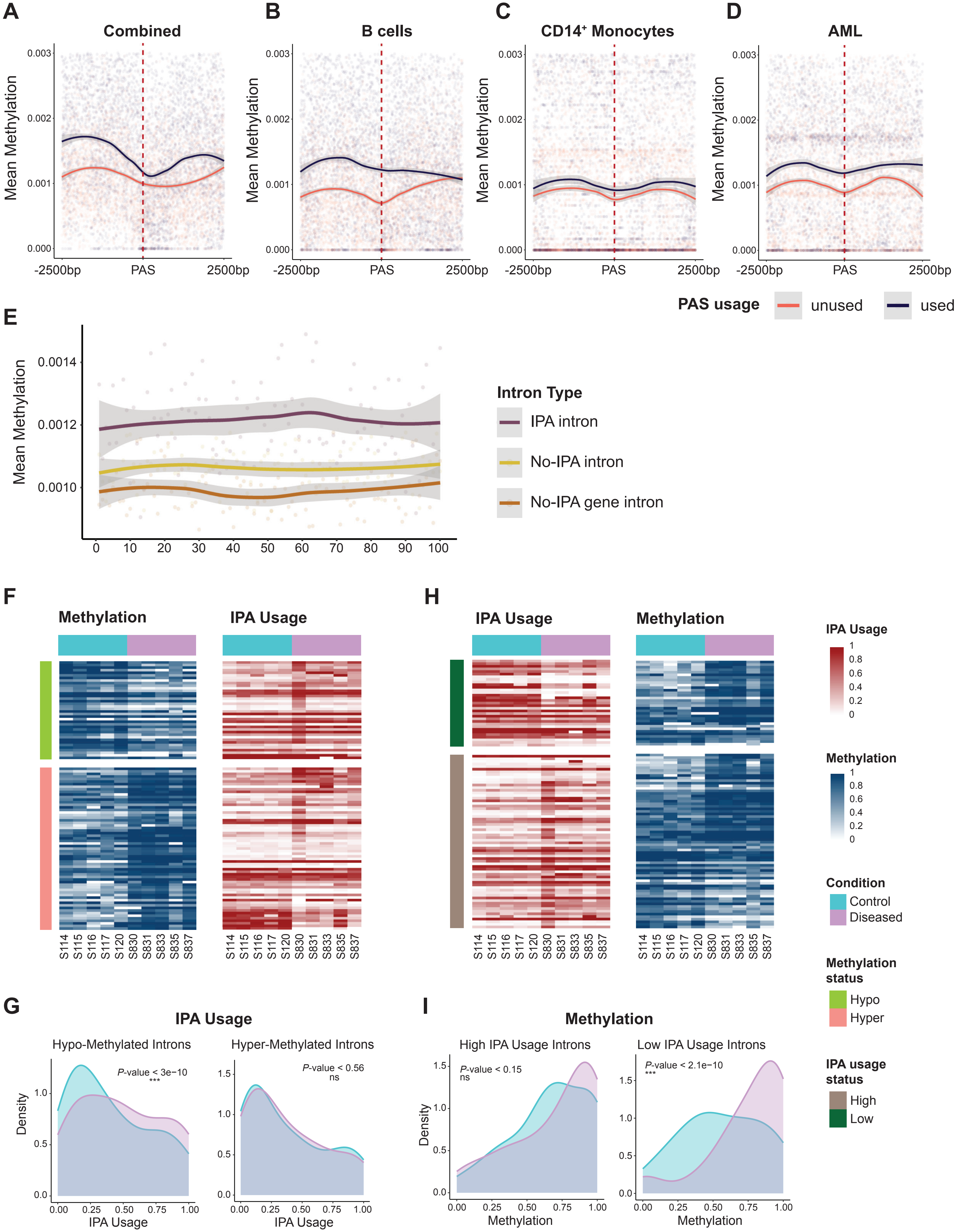

A
KDM4B (Chr19: 5075471−5084366(+))
Intron with Composite IPA
RNA-seq
60
40
20
0
30
20
10
0
3'-seq
30
20
10
0
1 kb
B
Penalty
4000
3000
2000
1000
0
Changepoints
0
1
2
C
Coverage
3
2
1
0
0
1000
2000
3000
4000
5000
Intron Start
Intron End
D
TPP2 (Chr13: 102602922−102616100(+))
Intron with Skipped IPA
RNA-seq
60
40
20
0
40
30
20
10
0
3'-seq
20
10
0
1 kb
E
Penalty
1200
1000
800
600
400
200
0
Changepoints
0
1
2
3
4
5
6
F
Coverage
8
6
4
2
0
0
2000
4000
6000
8000
Intron Start
Intron End


A
Composite IPA
ING5 (Chr2: 241709072−241715429)
RNA-seq
100
0
100
0
3'-seq
60
40
20
0
1 kb
B
Skipped IPA
EXOC4 (Chr7: 133252937−133254413)
RACE Primers
RNA-seq
80
60
40
20
0
184bp
30
20
10
0
3'-seq
12
8
4
0
0.1 kb
C
10kb
3kb
1kb
0.5kb
0.1kb
Ladder
EXOC4


A
Precision
Recall
F1-Score
Value
0.0
0.1
0.2
0.3
Test Sample 2
Test Sample 1
Test Sample 8
Test Sample 4
Test Sample 3
Test Sample 14
Method
IPAseek
IPAFinder
InPACT
B
IPA Expression ($\log_2$(IPA TPM + 1))
0
3
6
9
False Negatives
True Positives
C
IPAseek
IPAFinder
InPACT
No. of IPAs
0
500
1000
1500
Annotation Source
3'-seq Atlas
PolyAsite
Refseq
Ensembl
No Annotation
D
No. of PAS (%)
0
25
50
75
100
Hexamer
AAUAAA
AUUAAA
UAUAAA
AAUAAU
AAUACA
AAGAAA
AAUGAA
AAUAUA
AGUAAA
AACAAA
AUUAUA
ACUAAA
AUUACA
CAUAAA
AAUAAG
AAUAGA
GAUAAA
AACAAG
No Hexamer
E
Density
0.0000
0.0025
0.0050
0.0075
Distance (nt)
0
50
100
150
Sample
Test Sample 14
Test Sample 3
Test Sample 4
Test Sample 8
Test Sample 1
Test Sample 2
F
PARVB
(Chr22: 44023452−44025923)
RNA-seq
3'-seq
0
50
100
150
200
1 kb
G
CEACAM8
(Chr19: 42586622−42589409)
RNA-seq
3'-seq
0
50
100
150
200
250
0
4
8
1 kb


A
IPA usage
1
0.8
0.6
0.4
0.2
0
Status
Up
Down
Condition
0 min
30 min
90 min
180 min
360 min
1440 min
2880 min
4320 min
PEX13
ACOT8
B
ACOT8
RACE Primers
4320 min
2880 min
1440 min
360 min
180 min
90 min
30 min
0 min
0.8
0
1 kb
C
PEX13
RACE Primers
1.5
1 kb
D
Ladder
PEX13
Ladder
ACOT8
10kb
3kb
1kb
0.5kb
0.1kb
E
Undifferentiated
Undifferentiated (DMSO)
25nM_PMA
24hr
48hr
72hr
50 µm
F
ACOT8
IPA Usage (RNA-seq)
IPA Usage (qRT-PCR)
0.0
0.1
0.2
0.4
0.6
DMSO 96 hr
PMA 6 hr
PMA 12 hr
PMA 24 hr
PMA 48 hr
PMA 72 hr
PMA 96 hr
G
PEX13
IPA Usage (RNA-seq)
IPA Usage (qRT-PCR)
0.0
0.1
0.2
0.3
0.4
0.6


A
GSE184264
Cluster3
Cluster2
Cluster1
PC2
PC1
Cell Type
Naive B cells
Naive CD4⁺ T cells
Naive CD8⁺ T cells
Granulocytes
Monocytes
NK Cells
B
GSE114922
Cluster4
Cluster2
Cluster1
Cluster3
Cell Type
Granulocytic Precursor
Monocytic Precursor
Erythroid Precursor
CD34⁺ HSPCs
C
IPA Usage
Lineage
Myeloid
Lymphoid
Cell Type
Naive B cells
Naive CD4⁺ T cells
Naive CD8⁺ T cells
Granulocytes
Monocytes
NK Cells
Group
Myeloid−enriched
Lymphoid−enriched
D
IPA Usage
Cell Type
Granulocytic Precursor
Monocytic Precursor
Erythroid Precursor
CD34⁺ HSPCs
E
RPN1(Chr3:128626139−128632950)
CD34⁺ HSPCs
Erythroid Precursor
Granulocytic Precursor
Granulocytes
Monocytic Precursor
Monocytes
Naive B cells
Naive CD4⁺ T cells
Naive CD8⁺ T cells
NK Cells
Myeloid
Lymphoid
1 kb


A
Combined
B
B cells
C
CD14⁺ Monocytes
D
AML
Mean Methylation
0.003
0.002
0.001
0.000
-2500bp
PAS
2500bp
PAS usage
unused
used
E
Mean Methylation
0.0014
0.0012
0.0010
0
10
20
30
40
50
60
70
80
90
100
Intron Type
IPA intron
No-IPA intron
No-IPA gene intron
F
Methylation
IPA Usage
S114 S115 S116 S117 S120 S830 S831 S833 S835 S837
H
IPA Usage
Methylation
S114 S115 S116 S117 S120 S830 S831 S833 S835 S837
IPA Usage
1
0.8
0.6
0.4
0.2
0
Methylation
1
0.8
0.6
0.4
0.2
0
Condition
Control
Diseased
Methylation status
Hypo
Hyper
IPA usage status
High
Low
G
IPA Usage
Hypo-Methylated Introns
P-value < 3e−10
***
Hyper-Methylated Introns
P-value < 0.56
ns
Density
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
IPA Usage
I
Methylation
High IPA Usage Introns
P-value < 0.15
ns
Low IPA Usage Introns
P-value < 2.1e−10
***
Density
Methylation