Briefings in Bioinformatics, 2026, 27, bbag295
https://doi.org/10.1093/bib/bbag295
Published: 4 June 2026
Problem Solving Protocol

OXFORD

APAdeg enhances differentially expressed gene inference by leveraging site-specific signals in APA-seq data

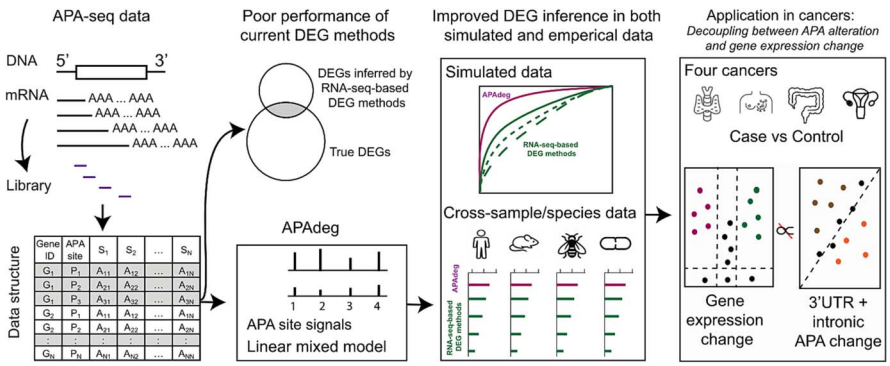
Bandhan Sarker[1], Tianjiao Zhou[2], Xiaoling Deng[1,2], Lei Zhang[1], Xianjia Zhao[1], Weijun Huang[2], Hongliang Yi[2], Hangjin Jiang[3,*], Chuan Xu[1,2,4,5,*]

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Otorhinolaryngology Head and Neck Surgery, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200233, China
[3]Center for Data Science, Zhejiang University, Hangzhou 310058, China
[4]Key Laboratory of Biodiversity and Environment on the Qinghai-Tibet Plateau, Ministry of Education, Xizang University, Lhasa 850000, China
[5]Lead contact
*Corresponding authors. Hangjin Jiang, Center for Data Science Zhejiang University, Hangzhou 310058 P. R. China. E-mail: jianghj@zju.edu.cn; Chuan Xu, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education Shanghai Jiao Tong University, Shanghai 200240, China, Phone: 86-(0)21-34207976. E-mail: chuanx@sjtu.edu.cn.
Biographical note As a Principal Investigator at Shanghai Jiao Tong University, China, Chuan Xu studies RNA diversity (e.g. APA), develops computational methods and explores its mechanisms, disease links and environmental adaptation roles.

Abstract

Alternative polyadenylation (APA) is a key post-transcriptional regulatory mechanism implicated in various diseases. Existing APA analysis tools are generally restricted to site detection and comparison, precluding differently expressed gene (DEG) analysis. Furthermore, standard RNA-seq-based DEG methods, though commonly used for gene expression profiling, demonstrate limited efficacy in identifying DEGs when directly applied to APA-seq datasets. To address this limitation, we developed APAdeg, a novel statistical method specifically tailored for DEG analysis of APA-seq data. APAdeg integrates both the total read count of a gene and the site-specific read counts within the gene into a generalized linear mixed model, thereby improving the efficiency of DEG detection. Benchmarking analyses on both simulated and empirical APA-seq data demonstrated that APAdeg consistently outperforms RNA-seq-based methods in DEG inference. Application of APAdeg to APA-seq data from distinct cancer types revealed that only a small proportion of DEGs exhibited significant changes in 3′ untranslated region length, with an equally small proportion showing significant alterations in intronic APA usage. To facilitate widespread adoption, we implemented APAdeg as an R package. Collectively, APAdeg significantly enhances the accuracy of DEG analysis from APA-seq data, thereby advancing research into APA-mediated gene regulation in diseases and health.

Graphical Abstract



Keywords alternative polyadenylation, differentially expressed gene, RNA-seq, APA-seq, new method

Received: March 12, 2026. **Revised:** May 3, 2026. **Accepted:** May 12, 2026

Introduction

Alternative polyadenylation (APA) is a widespread post-transcriptional regulatory mechanism in eukaryotes through which a single gene generates multiple mRNA isoforms bearing distinct 3′ ends. Recent genome-wide surveys revealed a high abundance of APA across multiple species [1, 2]. For example, ∼70% of human genes show APA, and ∼50% have three or more polyadenylation sites (PASs), underscoring its pervasive role in shaping transcriptomic diversity and gene regulatory complexity [3, 4]. By selecting different PASs, APA alters 3′ untranslated region (3′ UTR) length, coding potential, mRNA stability, nuclear export, and translation efficiency of genes [3, 4]. Consequently, the functional output of a gene may be substantially reshaped through APA-mediated regulation, enabling context-dependent or condition-specific roles during development, homeostasis, and disease [3, 5]. In particular, APA-mediated regulation at the RNA level can introduce or eliminate cis-regulatory elements within the 3′ UTR, leading to altered gene expression. For instance, in gene *Rras2*, APA-mediated 3′ UTR lengthening in senescent cells generates additional RBP-binding sites, ultimately reducing Rras2 protein abundance through splicing factor TRA2B-induced repression [6]. Conversely, through APA-mediated 3′ UTR shortening, *FNDC3B* in nasopharyngeal carcinoma removes pre-existing microRNA (miRNA)-binding motifs to enhance its expression, leading to elevated cell proliferation, migration, invasion, and metastasis [7]. Although such examples remain limited, they illustrate the substantial impact of APA on gene expression regulation and highlight the need for more accurate approaches to systematically identify APA-mediated regulatory mechanisms across diverse biological contexts.

To elucidate the relationship between APA dynamics and gene expression regulation, the first essential step is to accurately identify the changes of APA and gene expression between case and control conditions. Advances in high-throughput RNA sequencing have greatly facilitated such analyses, such that three strategies are commonly used for large-scale screening. First, RNA-seq, which is defined in this study as a short-read sequencing method that captures fragments spanning the entire RNA transcript, can reliably quantify differentially expressed genes (DEGs); however, APA changes can only be inferred computationally, and the predictive accuracy of these inference-based approaches remains limited (Figs 1a and S1A). Second, combining RNA-seq with APA-seq (a short-read sequencing method that specifically captures 3′-end RNA fragments) enables precise detection of both differential gene expression and APA shifts, but the requirement for two sequencing assays substantially increases experimental cost. Third, many studies rely solely on APA-seq, using it both to quantify APA changes and—by applying RNA-seq-based tools—to identify DEGs. In practice, however, repurposing RNA-seq-based algorithms for APA-seq data performs poorly. For example, in our benchmark dataset, DEGs identified by DESeq2 on APA-seq data showed substantial discrepancies compared with those derived from matched RNA-seq data (Figs 1b and S1B). These inconsistencies likely arise from the intrinsic differences between APA-seq and RNA-seq data characteristics, rendering DESeq2—an algorithm originally optimized for RNA-seq—suboptimal for DEG analysis using APA-seq data. Despite this limitation, most prior APA studies lacking matched RNA-seq data have nevertheless applied RNA-seq-based DEG tools directly to APA-seq datasets, which consequently introduces systematic biases and imposes substantial burdens on downstream functional validation. Therefore, establishing an APA-seq-appropriate method for quantifying DEGs is urgently needed.

Although numerous algorithms and software packages have been developed for APA research, they are primarily dedicated to the detection and comparison of APA sites across various data types [8]. For instance, tools such as APAlyzer [9], DaPars [10], and TAPAS [11] are designed for RNA-seq data, whereas PolyAseqTrap [12] and PolyA-miner [13] are tailored for APA-seq data. Additionally, scDaPars [14] and scAPAtrap [15] focus on APA site detection using 3′-end single-cell sequencing. Without exception, these methods focus on the characterization of APA site rather than the inference of DEGs from APA-seq data [8]. It is crucial to note that the analysis of APA changes primarily focuses on the detection and comparison of position and usage of APA site. In contrast, DEG analysis aims to quantify differences in gene expression levels (e.g. transcript abundance) at the gene level. Given these fundamental distinctions between the two analytical strategies, there remains a lack of specialized tools for inferring DEGs specifically from APA-seq data, despite the abundance of existing APA-related algorithms.

To address this issue, we developed APAdeg in this study, a novel statistical method specifically designed for the identification of DEGs from APA-seq data. By incorporating APA site-specific read counts while accounting for the total read counts within each gene, APAdeg employs a negative binomial (NB) generalized linear mixed model to effectively improve the accuracy of DEG detection. In both simulated and real APA-seq datasets, APAdeg outperforms conventional RNA-seq-based DEG methods. We applied APAdeg to case–control comparisons across multiple APA-seq datasets derived from four distinct cancers, systematically investigating the relationship between APA alterations and changes in gene expression levels in cancers. APAdeg has been implemented as an R/Bioconductor package to facilitate its broad adoption by the research community (https://github.com/XuLabSJTU/APAdeg).

Results

Specificity of APA-seq data compromises the efficacy of RNA-seq-based differently expressed gene methods

In RNA-seq analyses, sequencing data are typically consolidated into read count matrices, where the count value for each gene corresponds to the number of mapped RNA reads or fragments. Because count data exhibit an inherent mean–variance dependence, numerous methods based on generalized linear models (GLMs) have been developed to identify DEGs [16–23]. Early approaches relied on the Poisson distribution, which can capture basic gene expression variation across biological replicates [18, 20, 23]. However, Poisson models are limited in their ability to explain expression variability because they assume equality between the mean and variance, often resulting in inflated type I error rates [16, 19, 21]. To accommodate overdispersion, where the variance exceeds the mean, the NB distribution was subsequently introduced as an alternative to the Poisson model [19, 21, 24]. For NB distribution-based DEG analysis, DESeq2 [19] and edgeR [21] represent the most well-established and broadly applied computational tools in the field.

Despite their robust performance in DEG analysis of RNA-seq data, those RNA-seq-based tools exhibit suboptimal performance when applied to APA-seq data for DEG analyses. For example, we

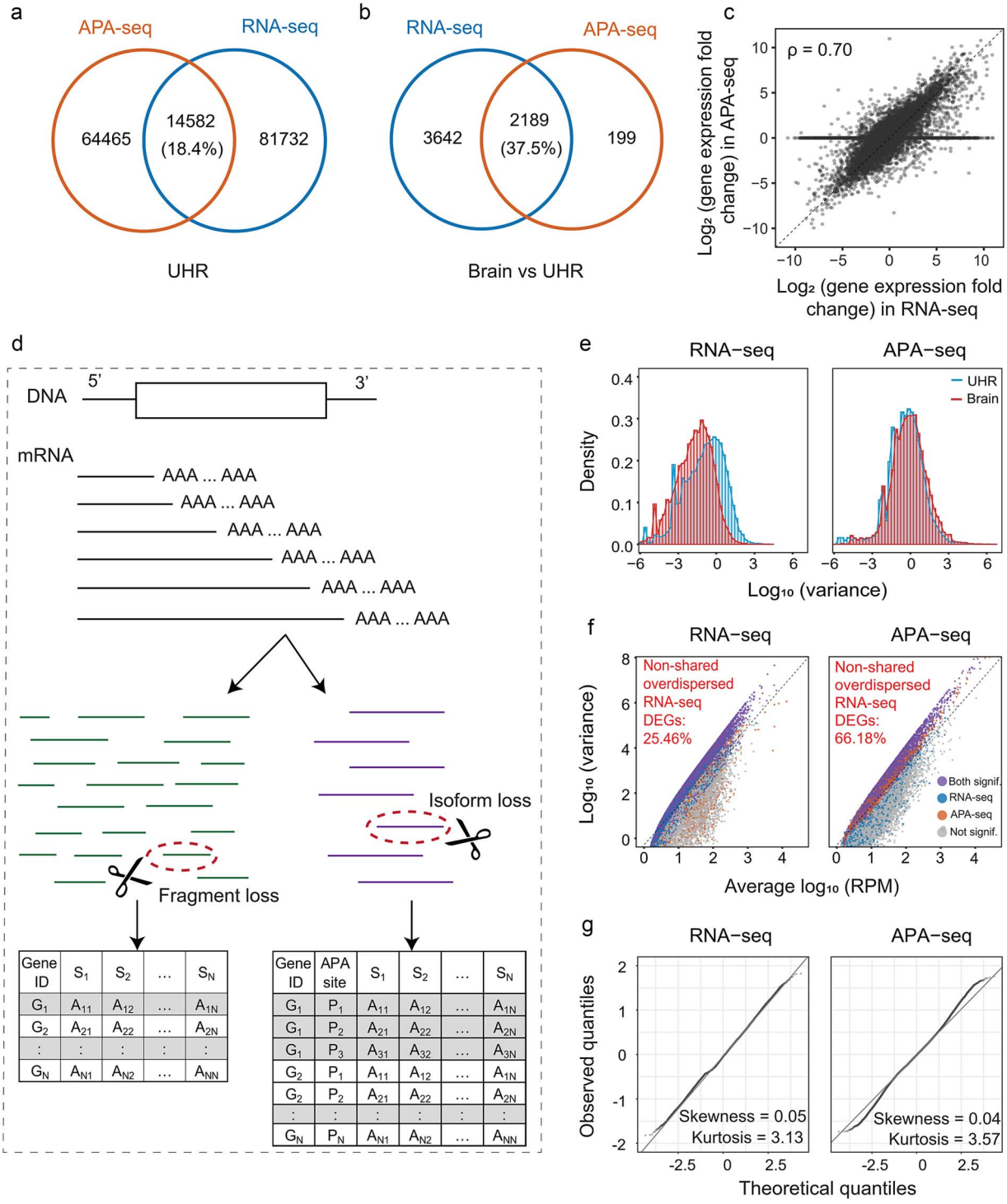


Figure 1 Specificity of APA-seq data compromises the efficacy of RNA-seq-based DEG methods. (a) Venn diagram illustrating the overlap between APA sites quantified by APA-seq and those inferred from RNA-seq in the UHR sample. The percentage shown indicate the proportion of overlapping sites relative to the total number of APA sites identified by APA-seq. (b) Venn diagram showing the overlap between DEGs detected by RNA-seq and those identified by APA-seq in the comparison of UHR versus brain samples. The percentages shown represent the proportion of overlapping genes relative to the total number of DEGs detected by RNA-seq. (c) Correlation between gene expression fold changes calculated from RNA-seq and those from APA-seq in the comparison of UHR versus brain samples. (d) Differences in library construction strategies and data structures between RNA-seq and APA-seq. (e) Distributional difference of variance between RNA-seq (left) and APA-seq (right) in the comparison between UHR and brain samples. (f) Scatter plot of mean–variance relationship between RNA-seq (left) and APA-seq (right). Among the non-shared RNA-seq DEGs identified by DESeq2, ∼25% and 66% exhibited overdispersion (variance exceeding the mean) in RNA-seq and APA-seq analyses, respectively. Genes are colored by significant status (blue: APA-seq only; red: RNA-seq only; purple: both; gray: not significant). (g) QQ plots of Pearson's residuals for RNA-seq (lef) and APA-seq (right).

downloaded publicly available paired RNA-seq and APA-seq datasets from human universal human reference (UHR) and brain samples. Then, DESeq2 was independently applied to each dataset to identify DEGs between UHR and brain. We designated the DEGs identified by RNA-seq as the reference standard, given its well-established reliability. Notably, only a small subset of DEGs detected by RNA-seq was recapitulated by APA-seq (37.5%; Fig. 1b). To further validate this observation, we performed RNA-seq and Quant-seq (a variant of APA-seq; see materials and methods) for a normal thyroid cell line (Nthy-ori 3-1) and two thyroid cancer cell lines (TPC-1 and BCPAP), respectively. DESeq2 was applied to identify DEGs between each cancer cell line and the normal counterpart. A consistent pattern was observed whereby only a small fraction of RNA-seq-derived DEGs could be recapitulated in APA-seq data (Fig. S1B). Furthermore, substantial discrepancies were detected between the gene expression fold changes calculated from RNA-seq and those derived from APA-seq datasets (Figs 1c and S1C). These observations demonstrate that RNA-seq-based DEG analytical pipelines are ill-suited for DEG analysis of APA-seq data.

To address why APA-seq data compromises the efficacy of RNA-seq-based DEG methods, we first analyzed the principle of APA-seq library construction. APA-seq encompasses a variety of sequencing strategies (e.g. Quant-seq, 3′-READS, and 3P-seq); regardless of the specific approach, the ultimate goal is to capture fragments corresponding to the 3′-end of RNA molecules. By contrast, RNA-seq is designed to capture fragments spanning the entire RNA transcript. While both RNA-seq and APA-seq are subject to technical biases, these biases act at fundamentally different levels (Fig. 1d). In RNA-seq, biases primarily affect fragment-level sampling and are averaged out across multiple fragments per RNA transcript, resulting in gene-level counts that remain proportional to transcript abundance. In contrast, APA-seq effectively samples each transcript once at its 3′ end, such that site-dependent capture efficiency directly determines whether an RNA molecule is counted, introducing condition-dependent bias into gene-level counts. These fundamental differences in library construction necessitate a detailed investigation into the distinct distribution characteristics of RNA-seq and APA-seq data.

To address this issue, we compared the distributions of gene read counts and their corresponding variances between RNA-seq and APA-seq datasets. We found that the distributions exhibited a unimodal pattern in both case and control groups (Fig. S1D). Notably, the distributions were very similar between case and control groups, regardless of whether RNA-seq or APA-seq data were analyzed (Fig. S1D). However, for RNA-seq data, the variance distribution of gene read counts differed largely between case and control groups; importantly, this intergroup variance difference was much larger than that observed in APA-seq data (Figs 1e and S1E). This result leads to substantial discrepancies in DEG detection between APA-seq and RNA-seq, which is consistent with our previous observations in Figs 1b and S1B.

To assess the suitability of the NB model underlying RNA-seq-based DEG methods (e.g. DESeq2) for APA-seq data, we first examined whether APA-seq satisfies the core NB assumption that variance exceeds the mean. Gene-level mean–variance analysis revealed that, whereas RNA-seq data largely followed the expected over-dispersed pattern, APA-seq data exhibited a substantial fraction of genes with variance lower than the mean, indicative of systematic under-dispersion, particularly among genes not detected as significant (Figs 1f and S2A). To evaluate the impact of this assumption violation on model fitting, we next examined quantile–quantile (QQ) plots of Pearson residuals from NB GLMs. In RNA-seq, residuals closely followed the theoretical normal distribution, indicating adequate model fit, whereas APA-seq showed pronounced and regular deviations characterized by an S-shaped QQ pattern with compression in the central region and extensive departures in both tails (Figs 1g and S2B). Together, these results indicate that APA-seq data deviate from the standard NB assumptions and harbor structured variability, i.e. not captured by NB-based models, thereby reducing statistical power and limiting the performance of RNA-seq-based DEG methods when applied to APA-seq data.

In summary, RNA-seq-based DEG methods perform sub-optimally on APA-seq data. This is mainly because APA-seq's specific library construction and data distribution poorly fit the assumptions of tools like DESeq2. Thus, APA-seq specificity compromises the efficacy of RNA-seq-based DEG methods, emphasizing the urgent need for novel algorithms leveraging this specificity to improve DEG identification.

Development of APAdeg tailored to differently expressed gene identification from APA-seq data

To develop a DEG analysis method specifically suited for APA-seq data, we first considered the unique data structure of APA-seq. In RNA-seq count matrices, rows represent genes and columns represent samples, with each entry corresponding to the total number of reads mapped to a given gene (Fig. 1d). In contrast, APA-seq data are organized with samples as columns but individual APA sites as rows (Fig. 1d). When RNA-seq-based DEG methods such as DESeq2 are applied to APA-seq data, read counts from multiple APA sites must be summed to obtain gene-level expression estimates, thus discarding APA site-specific information. This raises an important question: does the ignored information on APA sites contribute to accurate estimation of gene expression differences?

Previous studies have suggested a significant negative relationship between APA site usage patterns and gene expression levels [25]. To confirm this relationship, we calculated the Shannon diversity index for each gene using APA-seq data, which captures both the number of APA sites and their relative abundances, and independently quantified gene expression levels using RNA-seq data. Consistent with prior report [25], we observed a significant negative correlation between gene expression levels (RNA-seq-based) and APA diversity (APA-seq-based) in the UHR sample (Fig. 2a). This relationship was consistently observed across four additional samples with matched RNA-seq and APA-seq data (Fig. 2b). Moreover, we also found that larger differences in APA Shannon index were associated with larger differences in gene expression levels in the case versus control comparisons. For example, when comparing brain with UHR, differences in APA diversity were strongly and positively correlated with expression fold changes (Fig. 2c). Similar patterns were observed across the comparisons between cancers and normal cells (Fig. 2d). These results indicate that APA isoforms are closely linked to both gene expression levels and their differential expression, suggesting that APA site-specific information may contribute to the accuracy of DEG inference.

Motivated by the observations, we developed APAdeg, a DEG analysis framework specifically tailored to APA-seq data that explicitly incorporates APA site-specific information alongside gene-level read counts (Fig. 2e). The design of APAdeg consists of five key steps (Fig. 2e), which are briefly outlined below.

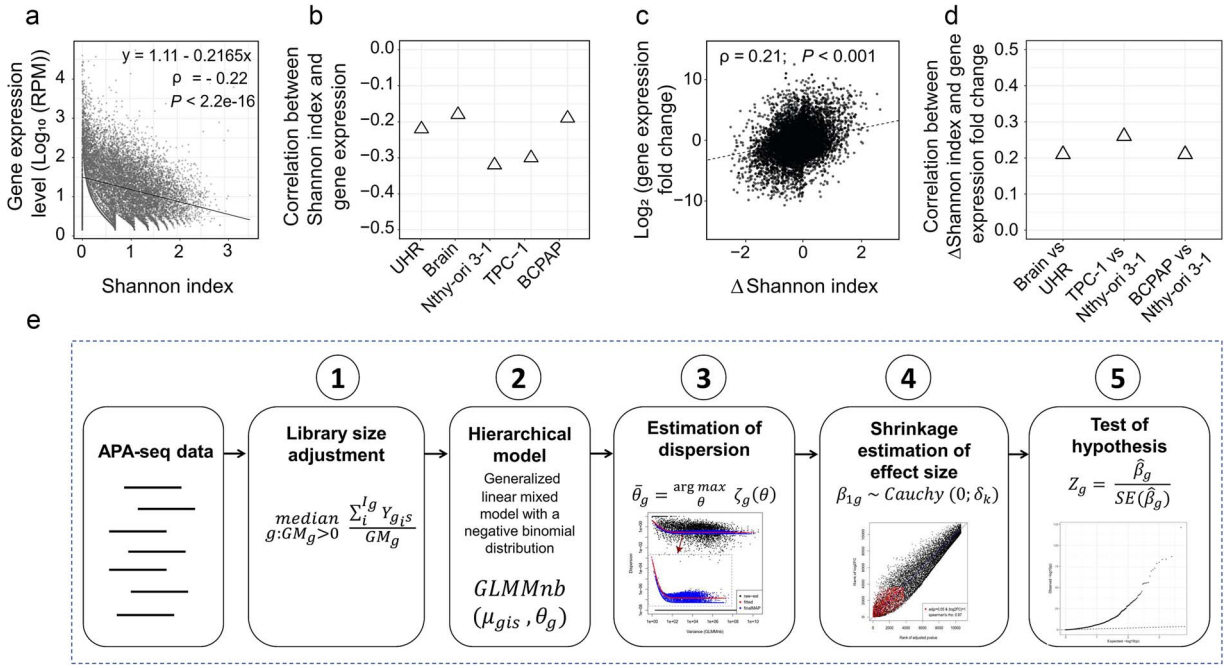


Figure 2 Development of APAdeg. Relationship between gene expression level and Shannon index of PAS diversity in UHR sample (a) and other four samples (b). Correlation between fold changes in gene expression and differences in PAS diversity in the comparison of UHR versus brain (c) and thyroid cancers versus normal cells (d). (e) Schematic overview of the workflow for the APAdeg model.

Step 1: APAdeg corrects for sample-specific sequencing depth and technical biases using median-of-ratios adjustment [19]. The resulting size factors are incorporated as offsets in downstream modeling, ensuring that observed differences in APA-seq read counts reflect biological variation rather than library size effects.

Step 2: adjusted APA-seq counts are modeled using a generalized linear mixed model with a negative binomial distribution (GLMMnb), in which APA isoforms are treated as hierarchical units nested within genes. This formulation explicitly accounts for overdispersion and isoform-level heterogeneity while encoding experimental conditions as fixed effects.

Step 3: to improve dispersion estimation, APAdeg fits gene-wise GLMMnb models and captures a global dispersion-variance trend across genes. An empirical Bayes shrinkage strategy is then applied to borrow information across genes, yielding stabilized dispersion estimates that enhance statistical power while controlling false positives.

Step 4: APAdeg applies adaptive shrinkage to log2-fold change (log2FC) estimates to stabilize effect sizes. This step attenuates exaggerated fold changes for lowly expressed genes and improves reproducibility, facilitating reliable gene ranking and biological interpretation.

Step 5: gene differential expression is assessed using Wald tests derived from the GLMMnb, with statistical significance evaluated independently of shrunken effect sizes. Multiple testing correction is applied to control the false discovery rate (FDR), resulting in DEG calls that are both statistically robust and biologically meaningful.

The above briefly summarizes the five steps of our algorithm development; full methodological details are provided in materials and methods. Additionally, each step involves relevant statistical models and processes, whose corresponding detailed content is separately included in the Supplementary Document S1.

APAdeg outperforms RNA-seq-based differently expressed gene methods on simulated APA-seq data

Simulation provides an ideal theoretical framework for evaluating the performance of newly developed computational methods. To assess the efficacy of APAdeg in identifying DEGs, we generated simulated RNA-seq and APA-seq datasets respectively based on real RNA-seq and APA-seq data (see materials and methods). Specially, we first randomly designated 10% of genes as true DEGs, with equal proportions of upregulated and downregulated genes (5% each). Subsequently, we generated simulated RNA-seq datasets (including both case and control conditions, with five biological replicates per condition) by mimicking the characteristics of real RNA-seq data derived from two biological replicates of human UHR and two biological replicates of human brain. Parallelly, we generated simulated APA-seq datasets (also encompassing case and control conditions with five replicates each) based on APA-seq data from the same set of human UHR and brain biological replicates.

We evaluated the performance of APAdeg and other comparison methods using four complementary metrics, including the area under the precision-recall curve (AUC-PR), precision-recall profiles, F-score, and FDR. We chose AUC-PR instead of receiver operating characteristic (ROC)-based metrics because DEG analysis typically involves highly imbalanced data, in which non-DEGs greatly outnumber DEGs. Under such conditions, precision-recall curves provide a more informative and reliable assessment than ROC curves [26, 27]. To minimize stochastic bias, the simulation procedure was repeated 50 times, and the average value of each performance metric was used for subsequent comparisons.

Although a variety of computational tools, such as APAlyzer [9], DaPars [10], and PolyAseqTrap [12], have been developed to

investigate APA, they are primarily designed for the detection of APA events rather than DEG inference. Therefore, we benchmarked APAdeg against nine widely used RNA-seq-based DEG methods, including DESeq2 [19], DESeq [16], edgeR [21], robust-edgeR [24], Brobustedger [28], limma-voom [29], SAMseq [30], NBPseq [31], and NOISeq [32], using the simulated APA-seq datasets with predefined DEGs. As shown in Fig. 3a, most RNA-seq-based DEG methods achieved moderate performance, with AUC-PR values ranging from 0.65 to 0.75, whereas NOISeq performed most poorly (AUC-PR < 0.5). In contrast, APAdeg achieved the highest AUC-PR (∼0.8). Consistently, the precision-recall curves (Fig. 3b) showed that APAdeg maintained higher precision across a broad range of recall values, indicating superior robustness under varying stringency thresholds. F-score analysis further supported this advantage (Fig. 3c). While most RNA-seq-based DEG methods yielded F-scores between 0.75 and 0.82 and NOISeq again showed poorest performance (F-score < 0.5), APAdeg achieved the highest F-score (∼0.9), reflecting a more favorable balance between precision and recall. Evaluation of FDR revealed that DESeq2, limma-voom, and SAMseq controlled FDR at acceptable levels (<0.1), whereas edgeR-based methods and NBPseq exhibited inflated FDRs (Fig. 3d). By comparison, APAdeg consistently achieved the lowest FDR (< 0.05), remaining well below the nominal threshold and indicating reliable control of false positives. Taken together, across all evaluated metrics, APAdeg consistently outperformed RNA-seq-based DEG methods on simulated APA-seq data, demonstrating superior accuracy, robustness, and false discovery control in DEG identification.

Because DEG sets derived from real RNA-seq data were used as reference standards in downstream analyses on real datasets, we adopted a consistent strategy in the simulation framework. Specifically, for each RNA-seq-based method, DEGs identified from simulated RNA-seq data were treated as the corresponding reference DEG set. We then quantified the concordance between DEGs identified from simulated APA-seq data and these RNA-seq-derived reference DEGs using the Jaccard index. The Jaccard index quantifies the similarity between two gene sets (calculated as the number of overlapping genes divided by the number of genes in the union of the two sets), with values closer to 1 indicating higher consistency. As shown in Fig. 3e, APAdeg always yielded a higher Jaccard index compared to all other methods, no matter which reference DEGs was used. For example, when using the DEGs identified by DESeq2 from simulated RNA-seq as the validation set, RNA-seq-based DEG methods produced Jaccard indices of less than 0.6, whereas APAdeg achieved a Jaccard index of >0.75 (top left subpanel in Fig. 3e). Additionally, when using consensus DEGs identified by all nine RNA-seq-based DEG methods from simulated RNA-seq data as the references, APAdeg recovered at least ∼10% more consensus DEGs than RNA-seq-based DEG methods (Fig. 3f), further confirming its superior ability to capture DEGs consistent with RNA-seq results.

We next investigated whether APAdeg's superior performance arises from reduced uncertainty in estimating group effects compared to RNA-seq-based DEG methods (Fig. S3). Using DEGs from simulated RNA-seq data as references, we compared the standard errors (SEs) of the estimated group effect (β) in simulated APA-seq data. APAdeg consistently produced smaller SEs for β across genes than RNA-seq-based methods (Fig. S3A), a pattern robust across benchmark RNA-seq methods. Across 50 simulation replicates, APAdeg reduced β's SE by ∼15% on average (Fig. S3B). For over 70% of reference DEGs, APAdeg yielded smaller SE estimates (Fig. S3C), indicating systematic uncertainty reduction. This translated into increased statistical power for detecting differential expression.

Together, these benchmarking results show that although RNA-seq-based DEG methods can achieve moderate performance on APA-seq data, they often fail to jointly optimize precision, recall, and false discovery control, and produce DEG sets with limited concordance to RNA-seq-derived results. In contrast, APAdeg reduces uncertainty in group effect estimation, and consistently outperforms all competing methods across multiple evaluation criteria, demonstrating higher sensitivity, improved reliability, and strong agreement with RNA-seq-based gene expression changes.

APAdeg exhibits superior performance in real APA-seq data with cross-sample/species applicability

Although APAdeg outperformed RNA-seq-based methods in DEG identification using simulated datasets, this does not directly validate its superior performance in real-world data. Therefore, we next evaluated the performance of APAdeg using real RNA-seq and APA-seq datasets. Initially, we conducted validation by comparing UHR and brain samples. Specifically, DEGs identified from RNA-seq data by each RNA-seq-based method were regarded as the gold standard (true DEGs) to assess the performance of APAdeg and nine RNA-seq-based methods in identifying DEGs from APA-seq data. The Jaccard index was retained as the evaluation metric, where a higher value indicates better predictive performance. For instance, DEGs of brain versus UHR calculated by DESeq2 were used as the gold standard (Fig. 4a, row 1). Subsequently, DEGs were computed from APA-seq data using the 10 aforementioned methods. Our results showed that the Jaccard index between DEGs identified by RNA-seq-based methods and the gold standard ranged from 0.21 to 0.40, which was obviously lower than that of APAdeg (0.46; Fig. 4a, row 1). When DEGs calculated by the other eight RNA-seq-based methods were separately used as the gold standard, APAdeg still achieved the highest Jaccard index values (Fig. 4a, rows 2–9). These results demonstrate that APAdeg consistently outperforms all other methods even when using real datasets.

Notably, the gold standard DEGs described above were derived from RNA-seq data by each respective RNA-seq-based method, which may introduce potential method bias. To eliminate such bias and enhance the reliability of the gold standard DEGs, we used the intersection of DEGs identified from RNA-seq data by the nine RNA-seq-based methods (Fig. S4) as the refined gold standard to re-evaluate the performance of all methods in identifying DEGs from APA-seq data. In the comparison between brain and UHR samples, the intersection of DEGs identified by the nine RNA-seq-based methods contained 1807 genes (Fig. S4A). Among these gold standard DEGs, 73.2% were successfully recovered by APAdeg from APA-seq data, which was significantly higher than the recovery rates of 0.6%–48.8% achieved by other methods (Fig. 4b). Thus, APAdeg still outperforms all other RNA-seq-based methods even when a more reliable set of gold standard DEGs is used.

To test the general applicability of APAdeg in APA-seq data from different samples and species, we further applied APAdeg to comparative analyses of additional datasets with both RNA-seq and APA-seq data, including human thyroid cancer cell lines (Fig. S5), mice, Drosophila, and *Schizosaccharomyces pombe* (Fig. 4). Regardless of whether DEGs calculated by individual RNA-seq-based methods from RNA-seq data were used as the gold standard or the intersection of DEGs from these

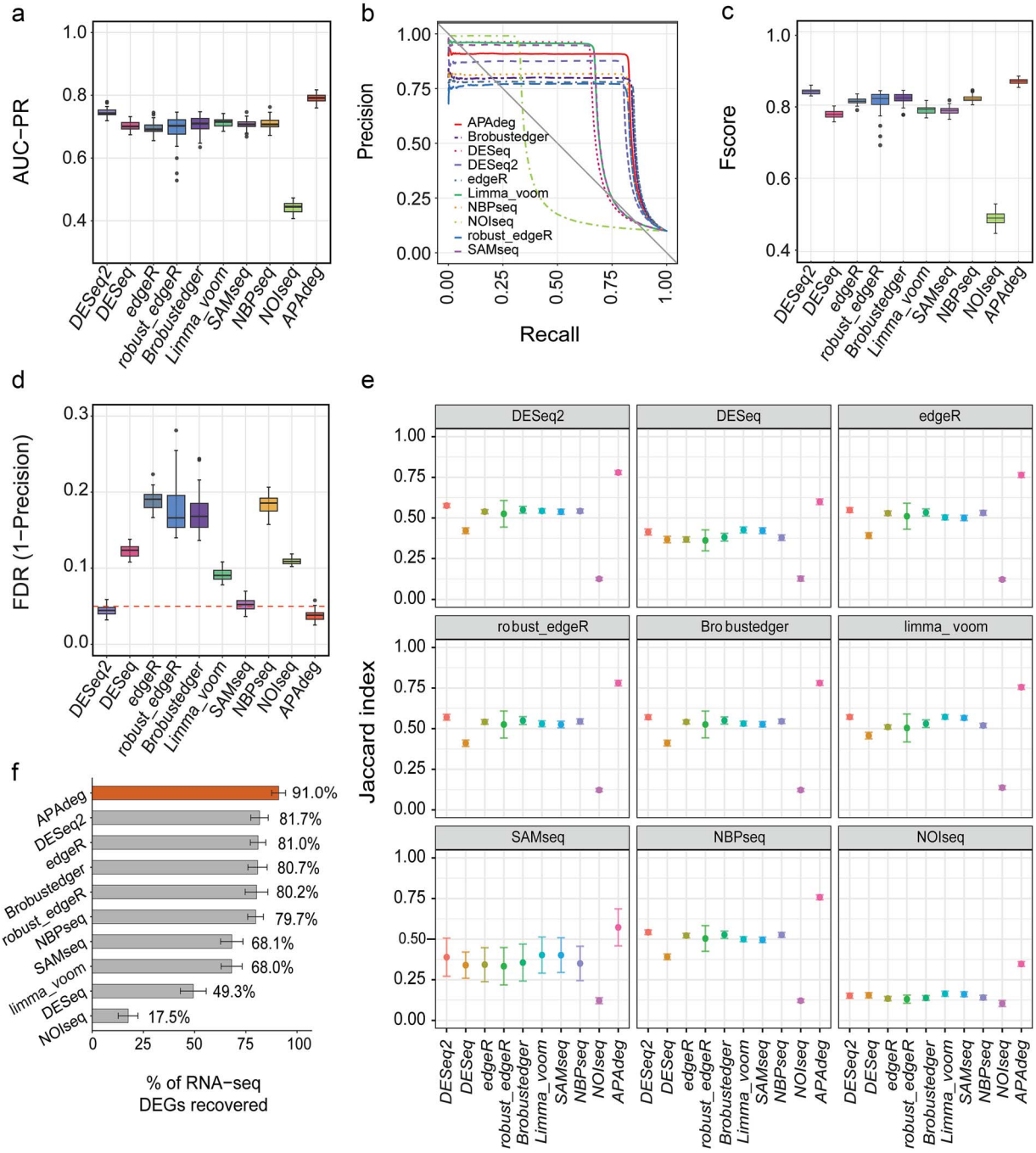


Figure 3 Evaluation of the performance of APAdeg through simulation study. Statistical performance of the proposed approach is evaluated. (a) Boxplots of area under the AUC-PR across simulated experiments. (b) Precision–recall curves showing average precision across recall levels. (c) Boxplots of the F-score, representing the balance between precision and recall. (d) Boxplots of FDR. (e) Jaccard indices quantifying the overlap between gene lists derived from APA-seq and RNA-seq across evaluation sets. DEGs in each subpanel are estimated from RNA-seq data based on respective methods. (f) Recovered percentage of consensus DEGs identified by all nine RNA-seq-based DEG methods.

methods was adopted as the more reliable gold standard, APAdeg consistently outperformed all other methods (Figs 4 and S5).

Collectively, these benchmarking revealed that while RNA-seq-based methods remain partially applicable to APA-seq data, their performance is limited across datasets and species. In contrast, APAdeg consistently achieved highest concordance with RNA-seq-derived DEGs and substantially improved recovery of reference DEGs. These findings highlight, APAdeg provides greater sensitivity, robustness and cross-species applicability for identifying DEGs from APA-seq data.

Gene expression changes are largely decoupled from 3′ UTR length alterations in cancers

APA has long been recognized as a regulatory mechanism of gene expression [3]. For instance, APA-mediated 3′ UTR lengthening in a

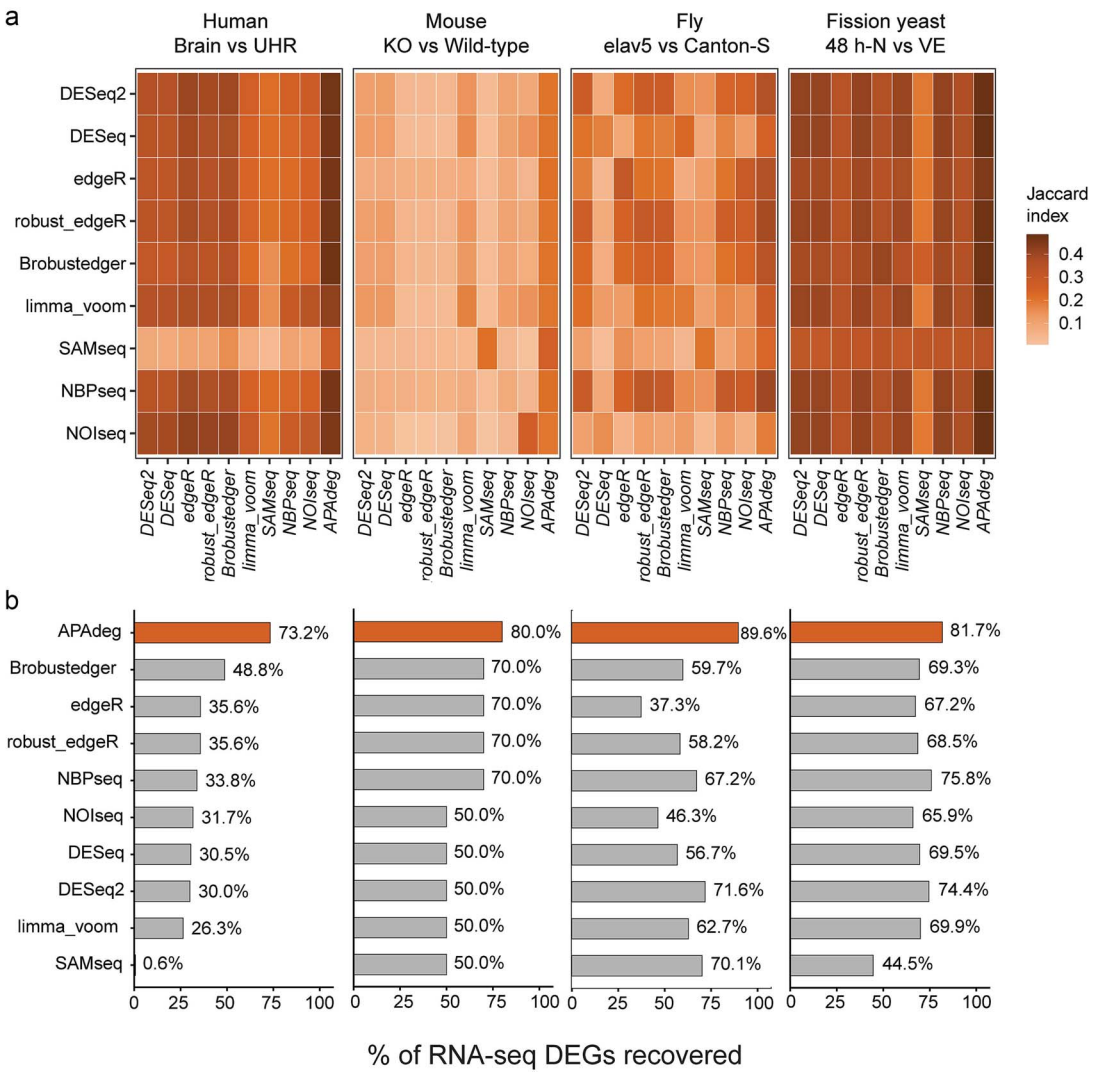


Figure 4 Comparison of *APAdeg* with other RNA-seq-based methods using real RNA-seq and APA-seq data. (a) Heatmaps displaying Jaccard index between DEGs identified by RNA-seq and those by APA-seq. Comparisons are shown for human (brain versus UHR), mouse (gene KO versus Wildtype), fly (elav5 versus CantonS), and fission yeast (48 h-N versus VE). The y-axis represents the reference DEGs identified from RNA-seq using corresponding methods, and x-axis represent the testing DEGs identified from APA-seq analyses using corresponding method. Color scale represents Jaccard index values. (b) Recovered percentage of consensus DEGs identified by all nine RNA-seq-based DEG methods.

specific gene can introduce additional miRNA binding sites within the extended region, thereby leading to miRNA-dependent downregulation of its transcript abundance. However, whether such APA-driven modulation of RNA levels through 3′ UTR remodeling constitutes a widespread phenomenon remains debatable. To address this question, we integrated DEG analysis with APA dynamics across multiple cancer models. In addition to our own APA-seq data from thyroid cancer cell lines (TPC-1 and BCPAP) and their normal counterpart (Nthy-ori 3-1), we compiled publicly available APA-seq datasets from three additional cancer types (incl., breast, colon, and ovarian cancers) along with their respective normal controls (see materials and methods). We first performed DEG analysis between TPC-1 and Nthy-ori 3-1 using APAdeg. This yielded 3865 DEGs (adjusted $P < .05$ and $|\log2FC| > 1$), comprising 2087 upregulated and 1778 downregulated transcripts (Fig. 5a). Comparable numbers of DEGs were identified in other cancer-normal comparisons (Fig. 5b).

To assess APA-associated 3′ UTR usage changes, we employed two complementary approaches. The first quantified the percentage of distal polyadenylation site usage (PDUI), defined as the ratio of reads mapping to the distal APA site over total reads spanning whole 3′ UTR region for a given gene (see materials and methods). A significantly higher PDUI in cancer versus control indicates preferential use of longer 3′ UTR isoforms in cancer, whereas a lower PDUI suggests 3′ UTR shortening. Using Fisher's exact test, we identified 2529 genes with significant 3′ UTR length alterations (D3UTRGs) in TPC-1 relative to Nthy-ori 3-1, including 1775 exhibiting 3′ UTR shortening (S3UTRGs) and 754 showing 3′ UTR lengthening (L3UTRGs; Fig. 5c). The second method applied the Mantel Haenszel test (MH-test) to directly compare 3′ UTR lengths inferred from sequencing coverage. MH-test uses both 3′ UTR length and the corresponding read count to assess statistical significance. APA alteration was quantified using UTR length index (ULI), which was estimated by Pearson correlation (see materials and methods). This approach detected 3978 D3UTRGs in TPC-1, comprising 2638 S3UTRGs and 1340 L3UTRGs (Fig. 5d). By intersecting the results from both methods, we defined a high-confidence set of 1423 S3UTRGs and 564 L3UTRGs in TPC-1 (Fig. 5e). Applying the same pipeline to

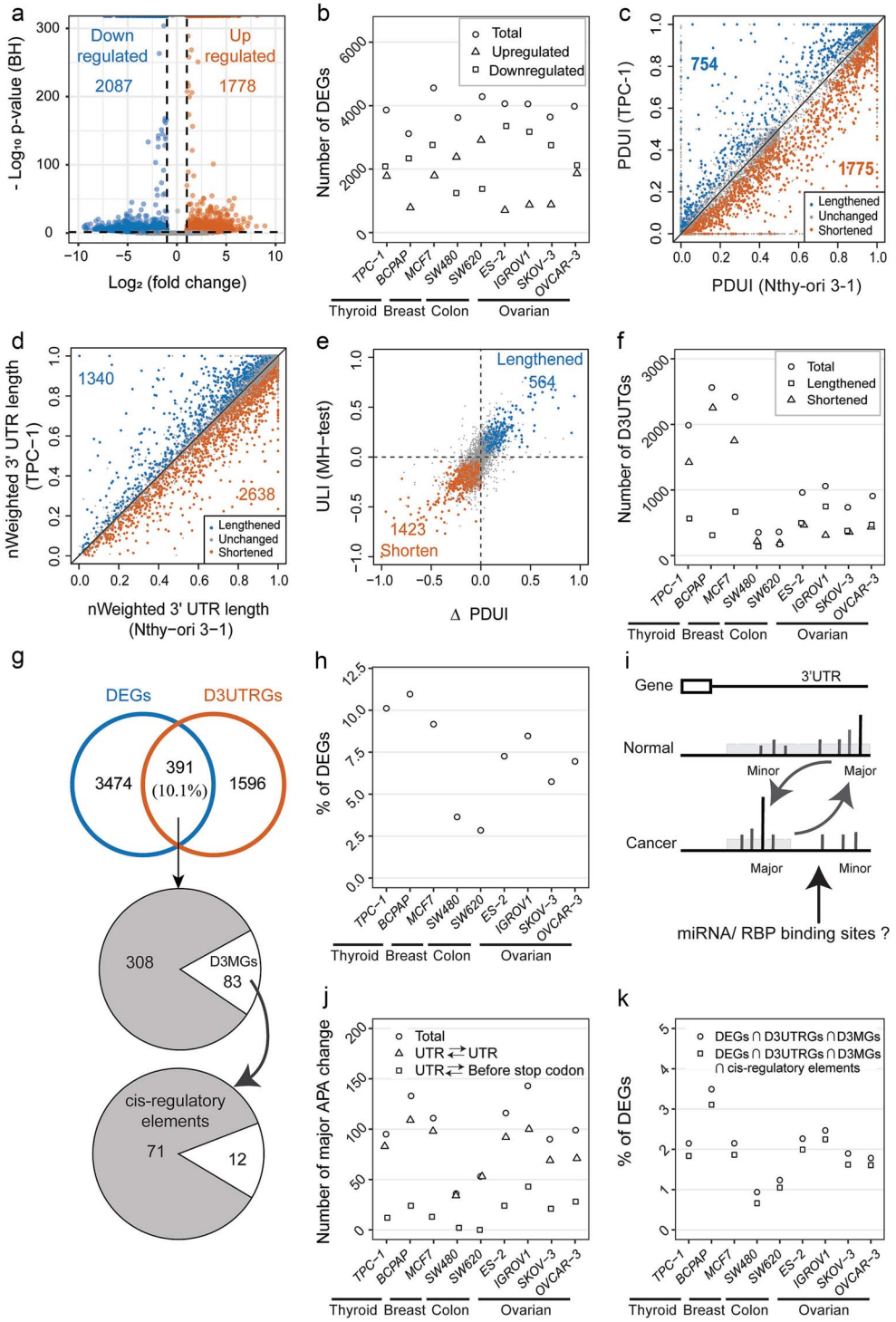


Figure 5 Correlation between alterations in gene expression and 3′ UTR length variation. (a) Number of DEGs in TPC-1 cells compared with Nthy-ori 3-1 cells. (b) Number of DEGs in four types of cancer cell lines relative to their corresponding normal cell lines. (c) Number of genes with differential PDUI values. (d) Number of genes with differential weighted 3′ UTR length. (e) Joint determination of 3′ UTR length variation by PDUI and weighted 3′ UTR length in genes. (f) Number of genes with 3′ UTR length variation in each of the four cancer types. (g) The extent of the potential association between gene expression and 3′ UTR length variation in TPC-1 versus Nthy-ori 3-1 cells. The top Venn diagram shows the overlap between DEGs and genes with 3′ UTR length variation; the middle pie chart shows the proportion of genes with altered major APA sites in the 3′ UTR among the above overlapping genes; the bottom pie chart shows the proportion of genes with differential cis-regulatory elements in the 3′ UTR resulting from the alteration of major APA sites among the above genes. (h) Overlap between DEGs and genes with 3′ UTR length variation in the four cancer types. (i) Schematic diagram of 3′ UTR alterations in genes, including the presence or absence of major APA site changes and the occurrence of cis-regulatory element variations in the altered regions. (j) Number of genes with altered major APA sites in the four cancer types. (k) Number of genes whose expression levels regulated by 3′ UTR length changes.

other cancers, we consistently found a predominance of S3UTRGs among D3UTRGs in thyroid, breast, colon cancers (Fig. 5f), supporting the well-established trend of global 3′ UTR shortening in cancers [33]. Interestingly, however, this 3′ UTR shortening pattern was not recapitulated in ovarian cancer, where genes exhibiting 3′ UTR lengthening were instead slightly more prevalent in ES-2, SKOV-3, and OVCAR-3, and markedly enriched in IGROV1 (Fig. 5f).

If APA-mediated 3′ UTR remodeling were a major driver of transcriptional dysregulation in cancer, a substantial overlap between DEGs and D3UTRGs would be expected. Contrary to this hypothesis, Venn analysis showed that only 391 genes overlapped between the DEG and D3UTRG sets in the TPC-1 versus Nthy-ori 3-1 comparison (Fig. 5g), representing merely 10.1% of all DEGs (Fig. 5h). Extending this analysis to eight additional cancer-normal comparisons yielded consistent results: the proportion of DEGs exhibiting concurrent 3′ UTR length changes never exceeded 11% (Fig. 5h). These findings indicate that, in the cancers examined, the majority of expression-altered genes do not undergo concomitant APA-mediated 3′ UTR remodeling, revealing a predominant decoupling between transcript abundance changes and APA dynamics.

Given our prior evolutionary analyses indicating that major APA sites (i.e. the highest expressed APA site) are typically functional, whereas minor APA sites are generally non-functional [25, 34], we hypothesized that functional expression changes are more likely driven by shifts in the major APA site (Fig. 5i). We therefore examined genes exhibiting differential usage of major APA sites (DMGs) between cancer and normal cells. Across nine comparisons of cancer-normal cells, we identified 36–143 DMGs, of which 34–109 involved shifts within the 3′ UTR (D3MGs), and 2–43 involved transitions between pre-stop-codon regions and 3′ UTR (DP3MGs; Fig. 5j). Given that we specifically aim to evaluate the relationship between 3′ UTR remodeling and changes in gene expression, we focused on D3MGs. We observed that only a small fraction of genes in the DEG ∩ D3UTRG intersection also qualified as D3MGs (Fig. 5k). For instance, in the comparison between TPC-1 and Nthy-ori 3-1, among the 391 genes that were both differentially expressed and exhibited altered 3′ UTR usage, only 83 displayed a switch in their major APA site within the 3′ UTR (Fig. 5g). These findings further indicate that, when considering functionally relevant APA events (i.e. changes in the major APA site), the association between APA dynamics and transcriptional dysregulation is even more limited.

Finally, even when both expression and 3′ UTR length are significantly altered, their co-occurrence may reflect independent responses to a common upstream perturbation rather than direct regulatory causality. To evaluate potential mechanistic connections, we analyzed the cis-regulatory elements of the altered 3′ UTR segments in genes that were simultaneously DEG, D3UTRG, and D3MG. Specifically, we predicted gains or losses of miRNA and RNA-binding protein (RBP) binding sites in the variable 3′ UTR regions (see materials and methods). Among the 83 genes belong to DEG ∩ D3UTRG ∩ D3MG in TPC-1, only 71 exhibited cis-element alterations attributable to APA-mediated sequence changes, representing just 1.84% of all DEGs (Fig. 5g). Similar low proportions were observed across other cancer types (Fig. 5k).

In summary, our integrative analyses across multiple cancers demonstrate that while APA is frequently altered in malignancy, its contribution to transcriptomic dysregulation via 3′ UTR-mediated mechanisms is limited. The majority of DEGs do not exhibit concordant APA changes, and even among those that do, some genes have no functional cis-regulatory rewiring. These results challenge the notion that APA is a primary driver of expression changes in cancers and instead support a model in which APA and transcriptional regulation operate largely independently.

Gene expression changes are largely independent from intronic alternative polyadenylation alterations in cancers

Intronic polyadenylation (IPA) represents another well-documented form of APA, characterized by premature transcription termination within intronic regions [3]. IPA events frequently produce mRNA isoforms that lack complete coding sequences, disrupt the canonical open reading frame, and often introduce premature termination codons. Such truncated transcripts are predominantly degraded via the nonsense-mediated decay (NMD) pathway, although NMD-independent decay mechanisms have also been reported [35], ultimately leading to reduced steady-state RNA levels. Consequently, IPA may serve as a post-transcriptional mechanism to downregulate gene expression—at least when expression is measured as total RNA abundance. However, the extent to which IPA contributes broadly to transcriptomic regulation across cancer contexts remains unclear. To address this issue, we systematically analyzed IPA dynamics in the same panel of cancer and normal cell lines described above. For each gene, we quantified the percent usage of IPA (PUIPA), defined as the ratio of reads supporting intronic APA sites to the total reads of whole gene. Comparing TPC-1 with its counterpart Nthy-ori 3-1, we identified 1380 genes exhibiting significant changes in PUIPA (DIPAGs), including 541 with increased and 839 with decreased IPA usage (Fig. 6a). Similar analyses across the other three cancer types revealed between 69 and 3375 DIPAGs per comparison, comprising 35–3129 genes with elevated and 34–839 with reduced IPA (Fig. 6b).

We next assessed the overlap between DIPAGs and DEGs. In the TPC-1 versus Nthy-ori 3-1 comparison, only 458 genes were both DEGs and DIPAGs, representing 11.8% of all DEGs (Fig. 6c). Across the other comparisons between cancer and normal cells, this proportion ranged from 0.63% to 19.98% (Fig. 6d), indicating that concordant changes in expression and IPA are rare.

Given that IPA typically triggers transcript destabilization [35, 36], any functional regulatory relationship between IPA and gene expression should exhibit directional consistency: increased IPA usage is expected to correlate with reduced RNA abundance, and vice versa. We therefore stratified DEGs into upregulated and downregulated subsets and DIPAGs into those with significantly increased or decreased IPA in cancers. We first examined the scenario in which upregulated DEGs are paired with decreased IPA. In the TPC-1 versus Nthy-ori 3-1 comparison, only 181 genes overlapped between these two categories, corresponding to 10.2% of upregulated DEGs (Fig. 6e). This proportion ranged from 0.34% to 10.2% across other cancer types (Fig. 6f). Conversely, we analyzed downregulated DEGs paired with increased IPA, a configuration most consistent with IPA-mediated repression. The overlap accounted for 0.29%–21.64% of downregulated DEGs before correction (Fig. 6g and h).

When both directional categories were combined, we estimate that only 0.33%–14.1% of all DEGs across the four cancer types exhibit IPA changes that are both statistically significant and directionally coherent with their expression alterations (Fig. 6i). Even when including genes whose expression could potentially be influenced by 3′ UTR

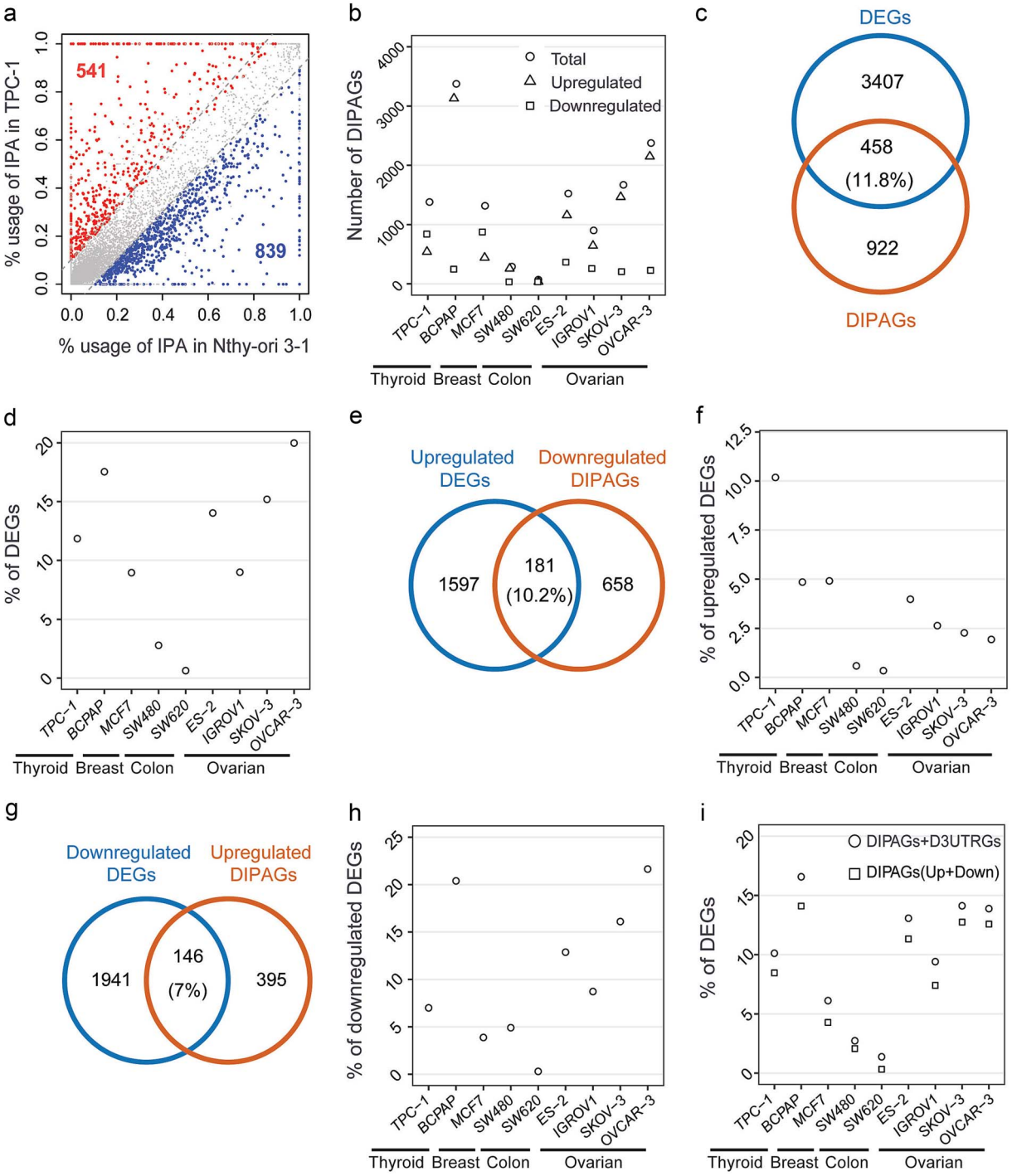


Figure 6 Correlation between alterations in gene expression and variations in intronic APA usage. (a) Number of genes exhibiting differential intronic APA usage in TPC-1 cells relative to Nthy-ori 3-1 cells. (b) Genes with differential intronic APA usage across four cancer types. (c) Overlap between DEGs and genes with DIPAGs in TPC-1 versus Nthy-ori 3-1 cells. (d) Extent of overlap between DIPAGs and DEGs across four cancer types. (e) Overlap between upregulated DEGs and downregulated DIPAGs in TPC-1 versus Nthy-ori 3-1 cells. (f) Overlap between upregulated DEGs and downregulated DIPAGs across four cancer types. (g) Overlap between downregulated DEGs and upregulated DIPAGs in TPC-1 versus Nthy-ori 3-1 cells. (h) Overlap between downregulated DEGs and upregulated DIPAGs across four cancer types. (i) Number of genes whose expression levels are potentially regulated by changes in intronic APA.

shortening or lengthening (Fig. 5k), this proportion remained limited at 1.38%–16.6% (Fig. 6i). Therefore, the majority of gene expression changes appear to occur independently of APA dynamics, suggesting that APA exerts only limited regulatory influence on RNA levels in these cancers.

Discussion

In APA research, investigating the interplay between APA dynamics and gene expression is of fundamental importance. While numerous software packages and algorithms have been established for APA

research, such as APAlyzer [9], DaPars [10], and PolyAseqTrap [12], these tools predominantly serve APA detection and comparison purposes, leaving a gap in dedicated methods for DEG analysis of APA-seq data. In the absence of matched standard RNA-seq data, a common practice in the field involves applying conventional RNA-seq-based tools, such as DESeq2 and edgeR, directly to APA-seq datasets for DEG identification. However, these methods are not optimally suited for APA-seq data (Fig. 1). To address this limitation, we present APAdeg in this study, a novel framework for DEG analysis specifically tailored for APA-seq data. The key innovation of APAdeg lies in its consideration of APA site-specific signals associating with gene expression (Fig. 2). Unlike conventional RNA-seq-based approaches that aggregate these signals into a single gene-level metric, APAdeg jointly incorporates both gene-level and site-specific read counts within a unified statistical framework. APAdeg effectively accommodates the inherent architecture of APA-seq data using GLMMnb and reduces uncertainty in group effect estimation, thereby substantially enhancing the accuracy of DEG inference. Benchmarking on both simulated and empirical datasets shows that APAdeg consistently outperforms RNA-seq-based DEG methods, positioning it as a superior tool for APA-seq analysis (Figs 3 and 4). Applying APAdeg to diverse cancers, we revealed that changes in gene expression are largely independent of alterations in 3′ UTR length and intronic APA usage (Figs 5 and 6), enabling a more precise dissection of post-transcriptional regulation in cancers. APAdeg is implemented as an R/Bioconductor package to promote widespread adoption in the research community.

Although the NB model is widely used for count-based expression data, standard generalized linear models (GLMnb) rely solely on fixed effects and perform poorly for correlated signal structures inherent to APA-seq. In APA-seq, gene expression is often correlated with PAS heterogeneity (Fig. 2a–d), introducing within-gene correlation that deviates from the assumptions of a GLMnb (Fig. 1f). To address this, APAdeg integrates PAS information as a random effect, accommodating diverse within-gene correlation structures and boosting statistical power. This enables GLMMnb model to capture structured variability and decompose total variance into components arising from random effects and residual overdispersion. As a result, the GLMMnb provides a more flexible representation of APA-seq data and reduces unexplained dispersion compared to the GLMnb (Fig. S6). Shared dispersion across genes stabilizes variance estimation and controls type I error in standard RNA-seq pipelines, yet this property is rarely implemented in mixed-effects models [37]. APAdeg explicitly incorporates shared dispersion estimation into its GLMMnb to maintain rigorous type I error control.

Accurate effect size estimation is critical for reliable gene ranking in APA-seq, but estimates are often unstable under low read counts or limited biological replicates—common constraints in APA-seq experiments. While shrinkage methods like apeGLM have been developed for standard GLMs, they do not support mixed-effects models tailored to APA-seq. To resolve this, APAdeg introduces Bayesian shrinkage to stabilize mixed-model effect size estimates and reduce false sign or effect size errors (Fig. S11), supporting more robust downstream inference. Fixed universal read-count filters are poorly generalizable across data types: low counts in unique molecular identifier-based RNA-seq and APA-seq can reflect genuine biological regulation, whereas those in non-deduplicated RNA-seq often arise from amplification noise. Such rigid filtering thus requires dataset-specific tuning and may distort biological signals. The Bayesian shrinkage approach in APAdeg circumvents these issues, offering a more appropriate and transferable strategy for APA-seq gene ranking.

APA-seq represents a specialized sequencing strategy designed to capture RNA 3′-end fragments, distinguishing it fundamentally from standard RNA-seq, which profiles the entire transcriptome (Fig. 1d). Consequently, standard RNA-seq remains the gold standard for high-throughput DEG analysis. Although the APAdeg tool is capable of performing DEG calculations, it is explicitly tailored for APA-seq data rather than standard RNA-seq datasets. APA-seq data can be generated via a wide array of 3′-end enrichment protocols, including 3′-READS, 3P-seq, PAT-seq, IVT-SAPAS, SAPAS, Quant-seq, 2P-seq, PAS-seq, MAPS, A-seq, A-seq2, 3′ T-fill, PAC-seq, and PolyA-seq [8]. Moreover, APAdeg is not restricted to APA-seq, but is applicable to datasets which gene-level expression exhibits hierarchical structure, reflecting transcript architecture and structured variability. For example, CAGE-seq [38] data, which profile alternative transcription start sites (ATSs), share a highly similar data structure with APA-seq, as both technologies capture only a specific end of transcripts. Notably, ATS diversity has also been reported to be negatively correlated with gene expression levels [39]. These parallels suggest that the APAdeg framework may be readily adaptable to CAGE-seq data for improved DEG analysis. Additionally, other sequencing modalities, such as 3′-end-based single-cell RNA-seq and transcript-centric full-length RNA sequencing (e.g. Iso-Seq), exhibit data characteristics analogous to those of APA-seq. The principles underlying APAdeg may therefore be extended to these platforms, underscoring the broader applicability of our methodological approach across multiple transcriptomic domains. Finally, APAdeg developed within a generalized linear mixed model framework, which is well-suited for cohort-scale APA studies [40], as they account for cohort heterogeneity, batch effects, and intra-sample variation. While the method outperforms conventional RNA-seq tools, achieving parity with standard RNA-seq quantification remains challenging due to intrinsic limitations in APA site-level expression measurement. Future work integrating refined bias correction and expression inference will further improve APAdeg's accuracy and applicability.

Applying APAdeg to multiple cancer types, we found that gene differential expression in cancer is largely decoupled from APA alterations, including changes in 3′ UTR length and intronic APA usage (Figs 5 and 6). Given the known differences between APA-seq-based and RNA-seq-based DEGs identification, we further evaluated this observation using available matched RNA-seq datasets with similar analytical techniques. In thyroid cancer, where APA-seq and RNA-seq data were generated from the same samples and experimental conditions, the largely decoupling between APA alterations encompassing both 3′ UTR length changes and intronic APA changes was robustly recapitulated (Figs S7 and S8). A similar trend was also observed in breast cancer, although RNA-seq data were obtained from the same cell line but different sources. These results support that observed decoupling is not solely a consequence of APA-seq-based DEG estimation, but may reflect a considerable biological phenomenon. Nevertheless, due to the availability of fully matched APA-seq and RNA-seq datasets remains limited, and further validation in additional cohorts would strengthen this conclusion. Furthermore, it is crucial to note that gene expression in this study is quantified by transcript abundance. Therefore, strictly speaking, the term "decoupling" specifically refers to the dissociation between transcript abundance and APA.

Several considerations are important for interpreting the decoupling between gene expression level and APA. First, even when APA-induced UTR changes introduce or remove cis-regulatory elements,

corresponding effects on gene expression may not manifest unless the relevant trans-acting factors are present in the same cellular context. Consequently, regulatory relationships inferred solely from cis-element annotations are inherently conservative, and the true extent of functional coupling may be even more limited. Second, the DEG analyses performed here are based on RNA concentration measurements. Although the overlap between genes exhibiting differential APA and differential expression level is small, this primarily reflects limited coupling at the RNA level. Gene expression can also be regulated at the protein level, and whether APA influences protein abundance in cancer remains an open question that warrants future investigation using matched proteomic data. Third, beyond expression levels, genes possess many additional functional attributes, such as subcellular localization and protein function, that may be affected by APA changes. Elucidating these alternative functional consequences represents an important direction for future studies.

Our findings of limited coupling between APA and gene expression in cancers are consistent with previous observations in proliferating T cells, where APA shortening was shown to have minimal effects on mRNA concentration [41]. Although that study relied on A-seq (a type of APA-seq) data and DEG inference using DEGseq, and therefore employed a less robust DEG framework than APAdeg (Figs 3 and 4), it arrived at a qualitatively similar conclusion. Moreover, several studies have reported weak or negligible effects of APA on protein abundance [41, 42], further supporting the notion that APA-mediated regulation of gene expression is not widespread. Together with our previous evolutionary analyses demonstrating that APA events are predominantly subject to negative selection [25], these results suggest that APA is more likely to represent a pervasive molecular byproduct rather than a broadly functional regulatory mechanism.

In summary, we developed APAdeg, a novel DEG inference framework specifically designed for APA-seq data. By jointly modeling gene-level and APA site-specific read counts within a mixed-effects and Bayesian statistical framework, APAdeg consistently outperforms conventional RNA-seq-based methods on both simulated and real APA-seq datasets. This approach enables more accurate dissection of the relationship between APA and gene expression regulation. Furthermore, the conceptual principles underlying APAdeg are readily extendable to other sequencing modalities with APA-like data structures, offering a generalizable strategy to improve differential analysis across diverse transcriptomic and multi-omic datasets.

Materials and methods

Real APA-seq and RNA-seq data

In this study, we acquired matched RNA-seq and APA-seq datasets across multiple model organisms, including *Homo sapiens* (encompassing cell lines and primary tissues), *Mus musculus*, *Drosophila melanogaster*, and *Schizosaccharomyces pombe*, for the construction and validation of our analytical model. Furthermore, APA-seq datasets derived from thyroid, breast, colon, and ovarian carcinomas were collected to evaluate the relationship between polyadenylation alterations and gene expression changes. A comprehensive inventory of all datasets utilized in this work is provided in Supplementary Table S1.

For comparative analyses between APAdeg and alternative computational methods, genes were defined as significantly regulated if they met the dual criteria of an adjusted *P*-value <.05 and an absolute $\log_2$-fold change (log2FC) > 1. The similarity between gene sets identified via APA-seq and RNA-seq was quantified using the Jaccard index, computed as: $Jaccard = \frac{A \cap B}{A \cup B}$, where A and B denote the gene sets inferred from APA-seq and RNA-seq profiling, respectively.

Simulation data

We generated simulated RNA-seq and APA-seq datasets independently using real empirical sequencing data as the reference. All simulation parameters, including gene-wise mean expression, dispersion, and the number of PAS per gene, were estimated from real RNA-seq and APA-seq profiles of two human UHR biological replicates and two human brain biological replicates [10]. Simulated read counts were sampled from an NB distribution with a mean μ_g and dispersion θ_g. We randomly designated 10% of all genes as true DEGs, with equal proportions of upregulated and downregulated genes (5% each). For both case and control conditions, we generated five biological replicates of simulated RNA-seq data by recapitulating the statistical characteristics of the real reference datasets; in parallel, matched simulated APA-seq datasets (five biological replicates per condition) were produced using the same set of human UHR and brain reference profiles. The RNA-seq-derived true DEGs served as the ground-truth gene set for evaluating APA-seq simulation analyses.

Differential expression testing was conducted, and *P*-values were adjusted for multiple testing using the Benjamini–Hochberg (BH) FDR procedure. Genes with an adjusted *P*-value <.05 were annotated as significantly differentially expressed. Model performance was quantified using the following metrics:

$Precision = \frac{True\ positives}{True\ positives+False\ positives}$, $Recall = \frac{True\ positive}{True\ positive+False\ negatives}$, and $Fscore = 2\ \frac{Precision \times Recall}{Precision+Recall}$. The AUC-PR was computed using the PRROC package in the R statistical environment.

APAdeg development

Step1: Library size adjustment

Raw APA site-level read counts are subject to sequencing biases arising from differences in sequencing depth across samples. To adjusting those library sizes, we estimated sample-specific size factors using the median-of-ratios method implemented in DESeq2 providing robust normalization by reducing the influence of highly expressed features [19]. Rather than normalizing counts prior to modeling, these size factors were incorporated into the GLMMnb as log-scale offsets (see Supplementary document S1), preserving the mean–variance relationship inherent to count data and enabling valid likelihood-based inference. This modeling strategy separates global sequencing effects from biological variation in APA usage while allowing the mixed-effects framework to capture gene- and site-level heterogeneity.

Step 2: Adoption of GLMMnb hierarchical model

APA-seq is a specialized form of RNA-seq but its data structure differs from RNA-seq. To identify DEGs for each gene g, multiple APA isoforms i are observed in each sample s (Fig. 1d). Let $Y_{g_i s}$ denote the uniquely mapped read counts for the isoform i of gene g in sample s (see Supplementary document S1).

In APA-seq data, a subset of genes exhibits expression in only one experimental condition, while being completely unexpressed across all samples in the opposite condition. This represents a non-trivial proportion (e.g. ∼15.07% in the brain versus UHR comparison; Fig. S9A) and is consistently observed across datasets ranging from 15% to 19% in the three examples of cell lines (Fig. S9B). Among these genes,

~17.5%–24.5% overlap with true DEGs identified from matched RNA-seq data (Figs S9C and S9D), highlighting their biological relevance. Directly estimating differential expression for such genes are statistically challenging through GLMM setting, as model parameters may lie on the boundary of the parameter space, leading to unstable fold-change estimates and inflated test statistics. To improve numerical robustness and parameter stability, APAdeg applies a regularization step by fitting a prior to the gene-wise GLMMnb under condition-specific non-expression. Specifically, expression levels in the non-expressed condition are regularized based on information from the expressed condition using a GLM. This procedure introduces a small, data-informed pseudo-expression that preserves the relative expression scale while avoiding degenerate likelihoods. As a result, APAdeg successfully recovers 62.2%–74.8% of those true DEGs in the three examined cell lines (Fig. S9E). Therefore, this regulatory step significantly improves the GLMM model complexity.

Following this preprocessing step, APA-seq counts are modeled using GLMMnb, which accommodating both overdispersion and hierarchical dependencies. Here, we assume that μ_{g_iS} is the expected count for isoform i of gene g in sample s, and θ_g is a gene-specific dispersion parameter. A random intercept b_{gi} incorporated to account for isoform-level variation across genes.

We further employ a log-linear GLMMnb design a matrix encoding sample-level covariates (e.g. treatment versus control status). The model yields an estimate of baseline isoform expression and $\log_2$-fold changes between groups. More generally, the GLMMnb framework allows the inclusion of additional random effects and interaction terms, thereby enabling analysis of complex experimental designs commonly encountered in APA-seq studies. This strategy allows APAdeg to robustly handle genes with extreme expression asymmetry while maintaining valid statistical inference.

Step 3: Information sharing across genes to improve differently expressed genes inference

In general RNA-seq data analysis, GLMs are widely used to estimate the relationship between mean and variance, parameterized as $\sigma^2 = \mu + \theta\mu^2$. In edgeR, the dispersion parameter θ is estimated from data and assumed constant across experimental conditions [16, 21]. Most of the methods developed for identifying DEGs agreed that sharing information across genes through data-driven relationships increases the power of detection [17, 19, 24]. Precise estimation of dispersion is critical in statistics, so GLM-based methods typically rely on a mean-dispersion trend based on the assumption that the mean expression adequately explains the variance across genes, where dispersion decreases as the mean increases. However, in our framework for APA-seq data using GLMMnb with variance $\sigma^2 = \mu + \frac{\mu^2}{\theta}$, yielding lower variance corresponds to larger θ. Variance of GLMMnb is influenced not only by the mean dispersion relationship but also by additional sources of variability such as APA site heterogeneity and random effects. As a result, genes with similar mean expression can exhibit substantially different variances. To better capture this heterogeneity, we hypothesized that sharing information across genes largely depends on variance rather than mean. Supporting this, we compared dispersion trends based on mean and variance trends. The mean-dispersion trend exhibited substantial over-smoothing, compressing dispersion estimates and showing a weak association between dispersion and mean (Fig. S10A). In contrast, the variance-dispersion trend better preserved the observed heterogeneity and showed improved agreement with empirical dispersion patterns (Fig. S10B). These observations motivated the development of a method to estimate dispersion from the dispersion-variance trend within the GLMM framework. Models are fitted for each gene and estimated gene-wise dispersion using restricted maximum likelihood (REML) or maximum likelihood (ML). We primarily fitted gene-wise GLMMnb with REML, which is less biased for variance component estimation. However, multiple correction tests failed to converge under REML, likely due to flat likelihood surfaces or boundary solutions. For these cases, we re-fitted the models using ML, which provided stable convergence while yielding comparable fixed-effect estimates. After that, we fitted a dispersion-variance trend based on a parametric/non-parametric formula (see Supplementary document S1). This fitted trend (red line in Fig. S10) represents the expected dispersion across genes but does not describe the variation of individual genes. To address this, we adopted an empirical Bayes shrinkage approach to shrink gene-wise dispersion similar to *DESeq2*. Furthermore, we estimated maximum posterior (MP) dispersion (blue dots) from the fitted trend. The shrinkage dispersion reduced potential false positives and improved power for detecting DEGs.

Step 4: Accurate estimation of effect sizes preserving true differences

Effect sizes are typically measured as log2FC, estimating true differences between experimental conditions, thereby enhancing gene ranking. However, estimating precise log2FC values for low-expression genes is challenging due to high variability (Fig. S11A, left). Such noisy genes distort ranking and downstream analyses. To overcome this, we applied an adaptive shrinkage approach based on empirical Bayes that stabilizes the estimator of log2FC (see Supplementary document S1). The yielding MP estimates attenuated exaggerated log2FC for low-expression genes (Fig. S11A, right). Shrunken log2FC values were more symmetrically distributed around zero and minimal deviation for weekly spread genes (Fig. S11B). Thus, shrunken log2FC reduced both biases and variance compared to row estimates, yielding a more reliable ranking.

Shrunken log2FC also provided more reproducible estimation than the raw estimated log2FC. To validate this, we used simulation data with six samples, and further split them into two equal groups (group I: 2 control versus 2 case A, group II: 2 control versus 2 case B). Between two balanced groups, we applied a fitted model to estimate raw log2FC (Fig. S12A) and shrunken MP to estimate shrunken log2FC (Fig. S12B) for each group. We observed that shrunken log2FC demonstrated greater resemblances between two groups, with a lower root mean square error (RMSE: 5.26) compared to raw log2FC (RMSE:11.65). These results supported that shrunken log2FC is consistent with experimental conditions, which may reduce bias estimation of effect sizes, preserving true biological differences.

Step 5: Hypothesis testing and differently expressed gene calling

The goal of the DEGs study is to estimate a set of genes that reject the null hypothesis of no expression difference between experimental conditions. The significant difference between the conditions is assessed by Wald test *P*-values. Love *et al.* initially assumed that shrunken log2FC divided by its SE could yield z-statistics for *P*-values under normal distribution [19]. However, the updated version of DESeq2 noted that shrunken log2FC should not be directly used for

hypothesis testing, as this violates the assumption of GLM [43]. Therefore, in our study, *P*-values were computed using the Wald test under GLMMnb, independent of shrunken log2FC (see Supplementary document S1). Multiple hypothesis testing procedure adopted for controlling FDR adjustment [44]. For example, QQ plots of UHR versus Brain samples showed that GLMMnb produced larger observed adjusted *P*-values than expected under the null, suggesting improved inference of estimating DEGs (Fig. S13).

There is a common phenomenon that genes that are significantly different even if the log2FC is very small or vice versa [45]. To evaluate the precision of shrunken log2FC, we compared adjusted *P*-values with log2FC estimates. A scatter plot of ranked adjusted p-values against ranked log2FC is shown in Fig. S14. We observed that rank-adjusted *P*-values against shrunken log2FC (Fig. S14B) are well calibrated, resulting in strong Spearman's correlation coefficients compared to raw log2FC (Fig. S14A). Together, these findings demonstrate that combining Wald test *P*-values with shrunken log2FC enhances both statistical power and biological interpretability in DEG detection.

Alternative polyadenylation alteration analyses

APA-seq data were processed using a standard analysis workflow. Briefly, raw reads were quality checked with FastQC [46], adapter trimmed using Cutadapt [47] and aligned to the reference genomes (hg38) with STAR [48] using default parameters. APA sites supported by fewer than two 3′-end-seq reads were removed and remaining sites within 24 bp were merged into initial clusters. If the initial cluster width > 24, those cluster are refined using weighted density peak clustering method [49]. For each cluster, the site with the highest read count was selected as the representative APA site, and the total read counts across the cluster were used as the site expression abundance. APA sites were assigned to genes if located between the gene 5′-end and 1000 bp downstream of the annotated 3′-end and expression levels were quantified as read per million. APA sites with the highest expression within a gene was defined as the major site. Predicted APA site from RNA-seq were inferred by TAPAS [11] pipeline with default parameters.

To identify the significant 3′ UTR APA change between conditions, we applied two complementary approaches: differential percentage of distal APA usage index (PDUI) and Mantel–Haenszel test (MH-test). Briefly, PDUI was estimated using the DaPars framework [10]. Because DaPars was originally designed for predicted APA events based on RNA-seq data, we adopted it to real APA sites by transferring site-level information into cumulative coverage profiles similar to RNA-seq coverage. For each gene, APA sites were ordered by genomic position and mean distal abundance (total distal coverage divided by distal length) and mean proximal abundance [(total proximal coverage minus mean distal abundance multiplied by proximal length) divided by proximal length] were calculated. A break point separating proximal and distal regions was identified by estimating a fit value for each APA site, which combined contributions from downstream and upstream sites based on genomic distance and differences in cumulative read abundance. The site with the lowest average fit value across conditions was selected as the breakpoint. Sites upstream of the breakpoint were defined as proximal, and downstream sites as distal. PDUI was calculated as the ratio of mean distal abundance to total abundance. Statistical significance was assessed using Fisher's exact test with BH correction. The change in PDUI (ΔPDUI) was defined as the difference between case and control conditions, where a negative ΔPDUI indicates 3′ UTR shortening and a positive ΔPDUI indicates 3′ UTR lengthening.

For MH-test was applied using two-ways contingency tables with ordered APA sites [50]. For each gene, UTR length were calculated based on the distance between poly(A) sites and the stop codon. Read counts for each APA site were arranged into contingency tables, with APA sites ordered from shortest to longest UTR length as columns and experimental conditions as rows. Pearson correlation used to calculate UTR length index (ULI) by taking the read counts as values and UTR lengths as column scores. Statistical significance between conditions was then evaluated using the MH-test with BH correction. A positive ULI indicates 3′UTR lengthening, while a negative ULI indicates 3′UTR shortening.

In addition, weighted 3′ UTR length for each gene was calculated as mean 3′ UTR length weighted by read counts of each 3′ UTR isoform. The normalized weighted 3′ UTR length (nWeighted 3′ UTR length) was defined as the isoform-weighted 3′ UTR length relative to the longest 3′ UTR length across the conditions [51].

To assess difference of intronic APA events, we calculated the percent usage of intronic APA sites for each gene as the ratio of intronic APA read counts to total APA read counts. We applied Fisher exact test with BH correction (*P*-value$<.05$) and $|\Delta$ percent usage$| < 0.1$ to identify significantly different usages of intronic APA between cancer and normal condition [52].

RNA binding protein/microRNA binding sites

miRNA target information was obtained from publicly available databases (https://www.mirbase.org/), and miRNA binding sites were predicted using the miRanda tool [53]. RBP binding sites were collected from curated databases (https://meme-suite.org/meme/db/motifs) and predicted using MEME tool with default parameters. High-confidence miRNA binding sites were defined using an energy cutoff < -20, while RBP binding sites were filtered using a q-value <0.05. Genes that losses or gains both miRNA and RBP binding motifs due to major APA changes were considered to harbor cis-regulatory elements in this study.

RNA-seq data processing

RNA-seq raw reads were quality-controlled using fastp [54] and aligned to the human reference genome (hg38) using STAR v2.7.10a [48]. Gene-level read counts for each sample were quantified using featureCounts [55] based on the STAR alignments.

R/Bioconductor package

APAdeg is implemented as a package for the R statistical environment. Development of APAdeg integrates NB function from MASS R package, mixed effect linear models from glmmTMB and lme4 R package and the empirical bayes approach from ashr R package. As input, it expects a data table where genes in rows have multiple isoforms and samples are in columns. Metadata and APAseq count data are stored in an S4 class derived from MetaData function within APAdeg R package. A single function called APAdegFit is used to run default analysis, while different parameters are also available for advanced users. Details of APAdeg R package implemented in vignette with pseudo examples for better understanding.

Key Points

- RNA-seq-based DEG analysis methods perform suboptimally when applied directly to APA-seq data.
- APAdeg uses a linear mixed-effects model leveraging site-specific signals to boost DEG detection in APA-seq.
- Benchmarking on simulated and empirical datasets confirms APAdeg consistently outperforms traditional RNA-seq-based methods.
- Application to cancer APA-seq data shows decoupling between gene expression changes and APA alterations, suggesting APA's limited role in transcript abundance regulation.
- APAdeg is available as an open-source R package to support research on APA-mediated gene regulation in health and disease.

Author contributions

C. X. conceived and designed the study. C. X. and B. S. wrote the paper. B. S. and C. X. assessed and validated the statistical approaches and models. B. S. and X. Z developed and released R package. H. J. contributed to the discussion of manuscript structure and focuses, and partially supervised the assessment of the statistical models. T. Z., L. Z., X. D., H. Y., and W. H. prepared partial data.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Funding

This study was supported by National Natural Science Foundation of China (32270704 and 32472630), National Science and Technology Innovation 2030 Major Projects for "Brain Science and Brain-Inspired Research" (2022ZD0214400), and Medical-Engineering Crossover Fund of Shanghai Jiao Tong University (YG2025QNB51).

Data availability

All datasets used in this study are provided in Supplementary Table S1. Specifically, the in-house generated thyroid cancer data and *S. pombe* data have been deposited in the National Genomics Data Center under the accession number PRJCA063560.

References

1. Hoque M, Ji Z, Zheng D. *et al.* Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods* 2013;**10**:133–9. https://doi.org/10.1038/nmeth.2288.
2. Derti A, Garrett-Engele P, Macisaac KD. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res* 2012;**22**:1173–83. https://doi.org/10.1101/gr.132563.111.
3. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* 2017;**18**:18–30. https://doi.org/10.1038/nrm.2016.116.
4. Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* 2013;**14**:496–506. https://doi.org/10.1038/nrg3482.
5. Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* 2019;**20**:599–614. https://doi.org/10.1038/s41576-019-0145-z.
6. Chen M, Lyu G, Han M. *et al.* 3′ UTR lengthening as a novel mechanism in regulating cellular senescence. *Genome Res* 2018;**28**:285–94. https://doi.org/10.1101/gr.224451.117.
7. Li YQ, Chen Y, Xu YF. *et al.* FNDC3B 3′-UTR shortening escapes from microRNA-mediated gene repression and promotes nasopharyngeal carcinoma progression. *Cancer Sci* 2020;**111**:1991–2003. https://doi.org/10.1111/cas.14394.
8. Zhang L, Deng X, Ma W. *et al.* Regulation and functions of alternative polyadenylation in fungi. *Mycology* 2025;**16**:1478–93. https://doi.org/10.1080/21501203.2025.2486776.
9. Wang R, Tian B. APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics* 2020;**36**:3907–9. https://doi.org/10.1093/bioinformatics/btaa266.
10. Xia Z, Donehower LA, Cooper TA. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat Commun* 2014;**5**:5274. https://doi.org/10.1038/ncomms6274.
11. Arefeen A, Liu J, Xiao X. *et al.* TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* 2018;**34**:2521–9. https://doi.org/10.1093/bioinformatics/bty110.
12. Ye W, Cheng X, Bi X. *et al.* PolyAseqTrap: a universal tool for genome-wide identification and quantification of polyadenylation sites from different 3′ end sequencing data. *Genome Biol* 2026;**27**:65. https://doi.org/10.1186/s13059-026-03963-w.
13. Yalamanchili HK, Alcott CE, Ji P. *et al.* PolyA-miner: accurate assessment of differential alternative poly-adenylation from 3′Seq data using vector projections and non-negative matrix factorization. *Nucleic Acids Res* 2020;**48**:e69. https://doi.org/10.1093/nar/gkaa398.
14. Gao Y, Li L, Amos CI. *et al.* Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res* 2021;**31**:1856–66. https://doi.org/10.1101/gr.271346.120.
15. Wu X, Liu T, Ye C. *et al.* scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Brief Bioinform* 2021;**22**:1–15. https://doi.org/10.1093/bib/bbaa273.
16. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106. https://doi.org/10.1186/gb-2010-11-10-r106.
17. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;**22**:2008–17. https://doi.org/10.1101/gr.133744.111.
18. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* 2010;**11**:R83. https://doi.org/10.1186/gb-2010-11-8-r83.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8.
20. Marioni JC, Mason CE, Mane SM. *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17. https://doi.org/10.1101/gr.079558.108.
21. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616.

22. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 2008;**9**:321–32. https://doi.org/10.1093/biostatistics/kxm030.
23. Wang L, Feng Z, Wang X. *et al.* DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 2010;**26**:136–8. https://doi.org/10.1093/bioinformatics/btp612.
24. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res* 2014;**42**:e91. https://doi.org/10.1093/nar/gku310.
25. Xu C, Zhang J. Alternative Polyadenylation of Mammalian Transcripts Is Generally Deleterious. *Not Adaptive, Cell Syst* 2018;**6**:734–742.e4. https://doi.org/10.1016/j.cels.2018.05.007.
26. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432. https://doi.org/10.1371/journal.pone.0118432.
27. Yang EW, Jiang T. SDEAP: a splice graph based differential transcript expression analysis tool for population data. *Bioinformatics* 2016;**32**:3593–602. https://doi.org/10.1093/bioinformatics/btw513.
28. Sarker B, Matiur Rahaman M, Alamin MH. *et al.* Boosting edgeR (Robust) by dealing with missing observations and gene-specific outliers in RNA-Seq profiles and its application to explore biomarker genes for diagnosis and therapies of ovarian cancer. *Genomics* 2024;**116**:110834. https://doi.org/10.1016/j.ygeno.2024.110834.
29. Law CW, Chen Y, Shi W. *et al.* voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29. https://doi.org/10.1186/gb-2014-15-2-r29.
30. Zhu A, Srivastava A, Ibrahim JG. *et al.* Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Res* 2019;**47**:e105. https://doi.org/10.1093/nar/gkz622.
31. Di Y, Schafer DW, Cumbie JS. *et al.* The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Stat Appl Genet Mol Biol* 2011;**10**:10. https://doi.org/10.2202/1544-6115.1637.
32. Tarazona S, Furió-Tarí P, Turrà D. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* 2015;**43**:e140. https://doi.org/10.1093/nar/gkv711.
33. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 2009;**138**:673–84. https://doi.org/10.1016/j.cell.2009.06.016.
34. Xu C, Zhang J. A different perspective on alternative cleavage and polyadenylation. *Nat Rev Genet* 2020;**21**:63–4. https://doi.org/10.1038/s41576-019-0198-z.
35. Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* 2007;**76**:51–74. https://doi.org/10.1146/annurev.biochem.76.050106.093909.
36. Berg MG, Singh LN, Younis I. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 2012;**150**:53–64. https://doi.org/10.1016/j.cell.2012.05.029.
37. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 2014;**11**:407–9. https://doi.org/10.1038/nmeth.2848.
38. Kodzius R, Kojima M, Nishiyori H. *et al.* CAGE: cap analysis of gene expression. *Nat Methods* 2006;**3**:211–22. https://doi.org/10.1038/nmeth0306-211.
39. Xu C, Park JK, Zhang J. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biol* 2019;**17**:e3000197. https://doi.org/10.1371/journal.pbio.3000197.
40. Rivellese F, Surace AEA, Goldmann K. *et al.* Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nat Med* 2022;**28**:1256–68. https://doi.org/10.1038/s41591-022-01789-0.
41. Gruber AR, Martin G, Muller P. *et al.* Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat Commun* 2014;**5**:5465. https://doi.org/10.1038/ncomms6465.
42. Spies N, Burge CB, Bartel DP. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res* 2013;**23**:2078–90. https://doi.org/10.1101/gr.156919.113.
43. Love MI, Huber W, Anders S. Moderation of estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
44. Benjamini Y, Drai D, Elmer G. *et al.* Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* 2001;**125**:279–84. https://doi.org/10.1016/S0166-4328(01)00297-2.
45. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* 2019;**35**:2084–92. https://doi.org/10.1093/bioinformatics/bty895.
46. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Cambridge, UK: Babraham Institute.
47. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* **17**:10–2. https://doi.org/10.14806/ej.17.1.200.
48. Dobin A, Davis CA, Schlesinger F. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635.
49. Ye C, Zhao D, Ye W. *et al.* QuantifyPoly(A): reshaping alternative polyadenylation landscapes of eukaryotes with weighted density peak clustering. *Brief Bioinform* 2021;**22**:1–14. https://doi.org/10.1093/bib/bbab268.
50. Fu Y, Sun Y, Li Y. *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* 2011;**21**:741–7. https://doi.org/10.1101/gr.115295.110.
51. Li Y, Sun Y, Fu Y. *et al.* Dynamic landscape of tandem 3' UTRs during zebrafish development. *Genome Res* 2012;**22**:1899–906. https://doi.org/10.1101/gr.128488.111.
52. Sun J, Kim JY, Jun S. *et al.* Dichotomous intronic polyadenylation profiles reveal multifaceted gene functions in the pan-cancer transcriptome. *Exp Mol Med* 2024;**56**:2145–61. https://doi.org/10.1038/s12276-024-01289-w.
53. Enright AJ, John B, Gaul U. *et al.* MicroRNA targets in Drosophila. *Genome Biol* 2003;**5**:R1. https://doi.org/10.1186/gb-2003-5-1-r1.
54. Chen S, Zhou Y, Chen Y. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560.
55. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656.

Briefings in Bioinformatics, 2026, **27**, bbag295
https://doi.org/10.1093/bib/bbag295
Problem Solving Protocol