## **Supplementary Method**

## **Detailed description of InPAS algorithms**

The InPAS workflow consists of four key steps including extraction of 3' UTR annotation, manipulation of coverage data, CPS search, and differential APA analysis. Additionally, InPAS provides functionalities for quality control of RNA-seq coverage data, visualization of differential APA events, and generation of files for functional analysis of APA events with GSEA.

### Extraction of 3' UTR annotation

InPAS streamlines the management of storage of metadata and intermediate files by setting up an SQLite database at the onset of the workflow. It selects the last exon of each annotated transcript as a proxy for the 3' UTR, from a user-provided TxDb database, which should correspond to the primary assembly of a reference genome to which the RNA-seq data of interest has been mapped. Next, representative non-overlapping segments are selected to remove any overlaps between 3' UTRs of transcripts on opposite strands. Then the 3' UTRs of the same gene are collapsed by their start coordinates, with a set of 3' UTRs of the same start coordinate represented by the longest one, and the end coordinates of shorter 3' UTRs stored as annotated CPSs for the representative 3' UTR. Any remaining representative 3' UTR not overlapping with others is extended to a user-defined distance or the closer boundary of the nearest downstream exon, whichever comes first. Each representative 3' UTR, along with its last segment of the coding sequence (CDS) (lastCDS) and any extended interval (gap), if available, is kept as GRanges. The complete set of representatives 3' UTRs is stored as a GenomicRangeList object, with one element per chromosome or scaffold (thereafter called chromosome).

### Manipulation of RNA-seq coverage data

InPAS first converts genome coverage data in bedGraph files into Rle (run length encoding) objects for efficient representation in R, with one Rle object per chromosome. The basewise coverage depth (total coverage depth) across the genome is summated and used to calculate depth weight later on. To enable chromosome-wise parallel computing, these Rle objects are grouped together by chromosome across all samples. For each chromosome, the following steps are performed for determining the distal and proximal CPSs: (1) Depth weight is computed by scaling the total coverage depth of each sample or experimental group to the mean total coverage depth across samples or groups, which normalizes for differences in base coverage depth; (2) The genome-wide coverage of each sample or group is summed by position to get total coverage. (3) Local background coverage is calculated for each extracted 3' UTR as the average coverage across intronic regions within a given range (1 kb, 5 kb, 10 kb, or 50 kb) centered on the transcript; (4) Coverage by group for each 3' UTR and associated gap is extracted from the results of Step 2. The 3' UTRs with mean coverage across the first 100 bases less than a coverage threshold (30× by default) in all groups are excluded for further analysis.

### Determination of CPSs

InPAS performs chromosome-wiseidentification of distal CPSs (dCPSs) and proximal CPSs (pCPSs) by analyzing the change pattern of RNA-seq read coverage over the 3' UTRs. To identify CPSs,, InPAS follows a four-step process, which involves searching for dCPSs, searching for pCPSs, adjusting the identified dCPSs, and adjusting the identified pCPSs.

To identify dCPSs, InPAS evaluates the change pattern of the normalized, merged coverage along each representative 3' UTR and the associated gap from 5' to 3' using a sliding window approach adopted from APAtrap [1].

Once dCPS search is complete, InPAS searches for pCPS using an optimal segmentation algorithm [2], if the distance between the start of 3' UTR and the inferred dCPS, , is greater than 10 bases or a user-defined cutoff. In this process, each base is a potential segmentation site, and a mean squared error (MSE) is computed for each base as follows:

where ,

is the number of experimental groups; is the length of the long 3' UTR in base; is the normalized read coverage of group *i* at base *j*; and are the estimated abundances of transcripts with long and hypothetic short 3' UTRs ending at *L* and *p* for experimental group *i*, respectively.

After computing the mean squared errors (MSEs) for each base *p*, the coordinate-ordered (from 5' to 3') MSEs for each 3' UTR are smoothed using the smooth.spline function from the R stats package. Locations with local minima of MSEs are identified as potential pCPSs and filtered to exclude those with *MSEmax* – *MSEj*≤ 0.1(*MSEmax* – *MSEmin*), where *j* is the location being considered. To determine the exact locations of the potential pCPSs, unsmoothed MSEs are used for local minima identification in the range of . Then, adjacent locations with local minima of unsmoothed MSEs are collapsed and represented by the location with the lowest local minima of MSEs, based on a user-defined maximal distance between two adjacent locations for collapsing. After collapsing, potential pCPSs that are not far away from the dCPS by a user-defined distance (200 bases by default) are excluded. The dCPSs can be adjusted by replacing them with a site of the highest probability among the downstream sequence of user-specified bases, identified by the naïve Bayes classifier from the cleanUpdTSeqpackage [3]. If a potential pCPS are co-localized with an annotated CPS within a user-defined distance, then the potential pCPS is replaced by the closest annotated CPS. Any remaining unannotated pCPSs can be adjusted using the same naïve Bayes classifier by searching the ± user-specified length of sequences flanking the predicted pCPSs.

### Differential APA analysis

InPAS is a versatile tool for differential APA analysisthat can be applied to both single and multiple samples from a variety of groups. It utilizes the Poisson hidden Markov model [4] for sample(s) from a single group, and incorporates Fisher’s Exact test for differential APA analysis between two groups without replicates, as adopted by DaPars [5]. Moreover, InPAS employs the powerful linear model and empirical Bayes statistical method using the diffSplice function from the limma package, which are designed to increase the statistical power for experiments with a limited number of replicates [6]. This is accomplished by borrowing information between genes to enhance the accuracy of the analysis. The first step is to calculate the percentage of distal poly(A) site usage index (PDUI) for each APA event. After calculating PDUI, InPAS log-transforms this data for use as input in the diffSplice function. The transformation ensures that the data adheres to the assumptions of the statistical model and allows for accurate and reliable identification of differential APA events between groups. PDUI is calculated as the ratio of distal site usage and total site usage for a given gene, as follows: , where and are the estimated abundances of transcripts with predicted long and short 3' UTRs ending at *L* and *p* for experimental group *i*, respectively [5].

Regardless of the statistical methods used, InPAS adjusts the nominal *p* values using the Benjamin-Hochberg procedure [7] to control the false discovery rate (FDR). When multiple potential pCPSs are found for a 3' UTR, InPAS only considers the one with the global minima for differential APA analysis.

### Other functionalities

InPAS can also output files for gene set enrichment analysis (GSEA) [8] and provide functions for genome coverage-based quality control of RNA-seq data and visualization of differential APA events.

## **References**

[1] Ye C, Long Y, Ji G, Li QQ, Wu X. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data.Bioinformatics*.* 2018; 34: 1841-1849.

[2] Lavielle M. Optimal segmentation of random processes.IEEE Transactions on Signal Processing*.* 1998; 46: 1365-1373.

[3] Sheppard S, Lawson ND, Zhu LJ. Accurate identification of polyadenylation sites from 3′ end deep sequencing using a naïve Bayes classifier.Bioinformatics*.* 2013; 29: 2564-2571.

[4] Lu J, Bushel PR. Dynamic expression of 3' UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling.Gene*.* 2013; 527: 616-623.

[5] Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ*, et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types.Nature Communications*.* 2014; 5: 5274.

[6] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies.Nucleic Acids Research*.* 2015; 43: e47-e47.

[7] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.* 1995; 57: 289-300.

[8] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis.Bioinformatics*.* 2007; 23: 3251-3253.