APA-Scan: Detection and Visualization of 3'-UTR APA with RNA-seq and 3'-end-seq Data

Naima Ahmed Fahmi[1], Jae-Woong Chang[2], Heba Nassereddeen[3], Khandakar Tanvir Ahmed[1], Deliang Fan[4], Jeongsik Yong[2,*], and Wei Zhang[1,*]

[1]Department of Computer Science, University of Central Florida
[2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Twin Cities
[3]Department of Electrical and Computer Engineering, University of Central Florida
[4]School of Electrical, Computer and Energy Engineering, Arizona State University
[*]To whom correspondence should be addressed.

February 2020

Abstract

The eukaryotic genome is capable of producing multiple isoforms from a gene by alternative polyadenylation (APA) during pre-mRNA processing. APA in the 3'-untranslated region (3'-UTR) of mRNA produces transcripts with shorter 3'-UTR. Often, 3'-UTR serves as a binding platform for microRNAs and RNA-binding proteins, which affect the fate of the mRNA transcript. Thus, 3'-UTR APA provides a means to regulate gene expression at the post-transcriptional level and is known to promote translation. Current bioinformatics pipelines have limited capability in profiling 3'-UTR APA events due to incomplete annotations and a low-resolution analyzing power: widely available bioinformatics pipelines do not reference actionable polyadenylation (cleavage) sites but simulate 3'-UTR APA only using RNA-seq read coverage, causing false positive identifications. To overcome these limitations, we developed APA-Scan, a robust program that identifies 3'-UTR APA events and visualizes the RNA-seq short-read coverage with gene annotations. APA-Scan utilizes either predicted or experimentally validated actionable polyadenylation signals as a reference for polyadenylation sites and calculates the quantity of long and short 3'-UTR transcripts in the RNA-seq data. The performance of APA-Scan was validated by qPCR.

Implementation: APA-Scan is implemented in Python. Source code and a comprehensive user's manual are freely available at https://github.com/compbiolabucf/APA-Scan

1 Introduction

Poly(A)-tails are added to pre-mRNA after the polyadenylation signal (PAS) during the 3'-end processing of pre-mRNA [5]. The last exon of mRNA contains a non-coding region, 3'-untranslated region (3'-UTR), which spans from the termination codon to the polyadenylation site. 3'-UTR is a molecular scaffold for binding to microRNAs and RNA-binding proteins and functions in regulatory gene expression [6]. In human and mouse, more than 70% of genes contain multiple PASs in their 3'-UTRs and APA using upstream PASs leads to the production of mRNA with shortened 3'-UTRs (UTR-APA) [1]. UTR-APA is known to increase the efficiency of translation and is associated with T-cell activation, oncogene activation, and poor prognosis in many cancers [4].

Several bioinformatics pipelines are available for the analysis of UTR-APA using RNA-seq data [8, 7, 2]. In general, all these methods measure the changes of 3'-UTR length by modeling the RNA-seq read density changes near the 3'-end of mRNAs. Indeed, with the aid of these methods, RNA-seq experiments became a powerful approach to investigate UTR-APA. In many cases, however, the identified APA sites are not functionally and physiologically relevant because most pipelines do not reference actionable PASs in their UTR-APA simulation. Moreover, none of the existing pipelines can provide high-resolution read coverage plots of the APA events with an accurate annotation. We have developed APA-Scan, a bioinformatics program for the detection and visualization of genome-wide UTR-APA events. APA-Scan integrates both 3'-end-seq (an RNA-seq method with a specific enrichment of 3'-ends of mRNA) data and the location information of predicted canonical PASs with RNA-seq data to improve the quantitative definition of genome-wide UTR-APA events. APA-Scan efficiently manages large-scale alignment files and generates a comprehensive report for UTR-APA events. It is also advantageous in producing high quality plots of APA events.

2 Methods

APA-Scan comprises of three steps: (i) read coverage estimation; (ii) identification of polyadenylation sites and the calculation of APA; (iii) graphic illustration of UTR-APA events (Figure 1). In the first step, APA-Scan takes aligned RNA-seq and 3'-end-seq data in the BAM format as an input to estimate the read coverage on 3'-UTR exons and identify potential polyadenylation sites. The read coverage files are generated by SAMtools [3]. In this step, the 3'-end-seq data is an optional input.

In the second step, all aligned reads from 3'-end-seq data are pooled together to identify peaks and the corresponding cleavage sites in 3'-UTRs, as shown in Figure 1. Identified peaks in the 3'-end-seq data are considered potential cleavage and polyadenylation sites. If the 3'-end-seq data is not provided by the user, predicted PASs (AATAAA, ATTAAA) in 3'-UTRs are considered as potential cleavage sites. Next, to determine potential 3'-UTR APA events

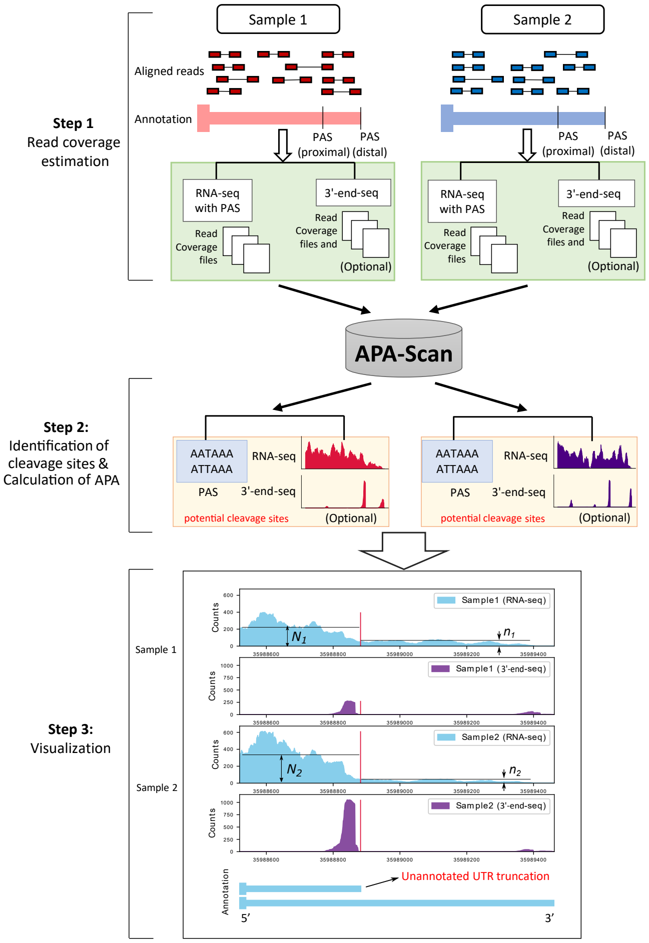


Figure 1: **Workflow of APA-Scan.** Starting with aligned RNA-seq and 3'-end-seq (optional) bam files, APA-Scan consists of three steps and generate high quality graphic illustration of aligned sequences with the indication of 3'-UTR APA events. The illustration also highlights unannotated short 3'-UTR transcript identified from this task. The vertical red lines show the corresponding cleavage sites.

between two biological contexts (or samples), APA-Scan evaluates each experimentally proven or predicted cleavage site in the 3'-UTR of a transcript using χ^2-test: it contrasts the RNA-seq short reads covering up and downstream of the candidate cleavage site between the two samples and calculates the mean coverage upstream of the site (N_1 and N_2) and downstream of the site (n_1 and n_2) as shown in at the bottom panel in Figure 1, with (N_1, n_1) denoting the coverage in the first sample, and (N_2, n_2) denoting the coverage in the second sample. Then, the canonical 2 x 2 χ^2-test is applied to report the p-value for each candidate site. All the identified events will be reported in an Excel file.

In the third step, based on the significance of 3'-UTR APA events calculated in the second step, APA-Scan can generate RNA-seq and 3'-end-seq (if provided) coverage plots with the 3'-UTR annotation for one or more user-specific events. In this step, users may specify the region of the genome locus to generate the read alignment plot. An example of this task is illustrated at the bottom panel of Figure 1.

3 Results

3.1 Experimental results

To validate the analysis results by APA-Scan, we conducted qPCR experiments for *Srsf3* and *Rpl22* transcripts from WT (wild type) and Tsc1-/- mouse embryonic fibroblasts (MEFs) based on the significant 3'-UTR APA events reported by APA-Scan. As shown in Supplementary Figure 1, both *Srsf3* and *Rpl22* showed the increase of the short 3'-UTR transcript by APA in Tsc1-/- compared to WT MEFs, which is consistent with our observations on the RNA-seq read coverage plots. These results further confirm that APA-Scan can identify the true 3'-UTR APA events with RNA-seq and 3'-end-seq samples from two different biological contexts.

3.2 Materials and Methods

Realtime quantitative PCR (RT-qPCR) analysis and primer sequences: Total RNAs from TSC1 WT or TSC1-/- MEF cells were isolated by Trizol method according to manufacturer's protocol `https://assets.thermofisher.com/TFS-Assets/LSG/manuals/trizol_reagent.pdf`.

Reverse transcription reaction using Oligo-d(T) priming and NxGen M-MuLV Reverse transcriptase (Lucigen) was carried out according to the manufacturer's protocol `https://www.lucigen.com/docs/manuals/MA115-M-MuLV.pdf`. SYBR Green was used to detect and quantitate the PCR products in real-time reactions. Quantitation of the real-time PCR results was done using standard curve method for accuracy and reliability of the analysis. The primer sequences used to measure the RSI for each transcript are as follows:
mRpl22 Total forward 5'-AAGTTCAC CCTGGACTGC AC-3'
mRpl22 Total reverse 5'-GTGATCTT GCTCTTGCTG CG-3'

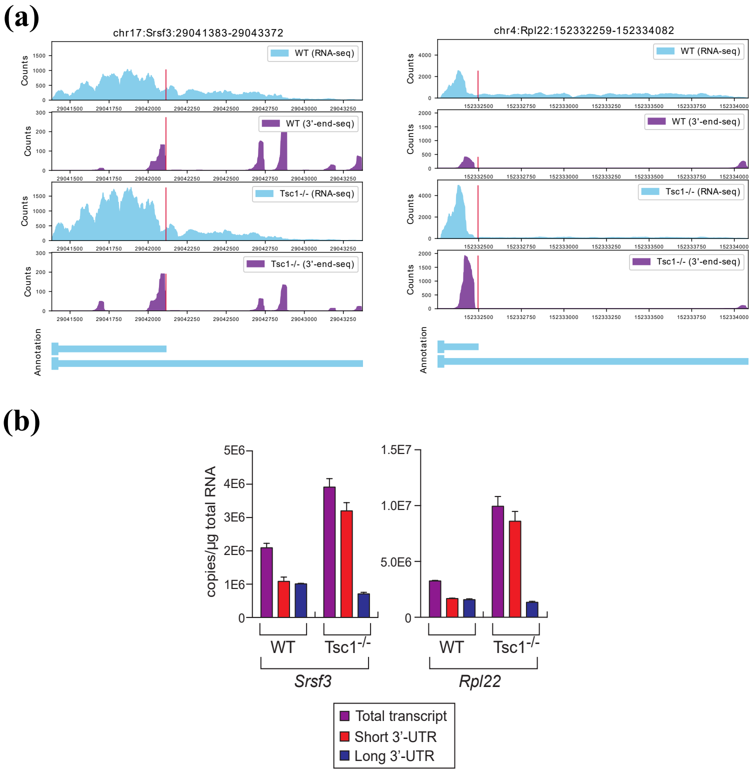


Figure 2: Experimental results. (a) RNA-seq and 3'-end-seq read coverage plots of the 3'-UTR in *Srsf3* and *Rpl22* gene in the two samples with isoform annotation. (b) The level of total, short 3'-UTR, and long 3'-UTR transcripts from *Srsf3* and *Rpl22* was measured by qPCR. Because it is not possible to design specific primers for the qPCR analysis of short 3'-UTR transcript, the amount of short 3'-UTR transcripts were calculated by subtracting the quantity of long 3'-UTR transcripts from total.

mRPL22 Long Forward 5'-TGGGCATC TGGGCTTTTA GG-3'
mRPL22 Long reverse 5'-GCTTGTTGCA GACTTGCTCA-3'
mSRSF3 Total forward 5'- GCTGCCGTGTAAGAGTGGAA-3'
mSRSF3 Total reverse 5'- AGGACTCCTCCTGCGGTAAT-3'
mSRSF3 Long forward 5'- TGCAACAGTCTTGTGGCTTA-3'
mSRSF3 Long reverse 5'-TGCAATGGCTCTTACATAGACC-3'

4 Conclusion

APA-Scan offers a computational pipeline to identify transcriptome-wide 3'-UTR APA events. By integrating RNA-seq data and PAS information (experimentally verified or computationally predicted), APA-Scan can generate a comprehensive report of significant APA events and the illustration of their read coverage plots. The wet-lab approaches using qPCR experiments demonstrate that APA-Scan provides high-accuracy and quantitative profiling of 3'-UTR APA events.

References

[1] Ran Elkon, Alejandro P Ugalde, and Reuven Agami. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, 14(7):496, 2013.

[2] Loredana Le Pera, Mariagiovanna Mazzapioda, and Anna Tramontano. 3USS: a web server for detecting alternative 3' UTRs from RNA-seq experiments. *Bioinformatics*, 31(11):1845–1847, 2015.

[3] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[4] Christine Mayr and David P Bartel. Widespread shortening of 3' UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, 2009.

[5] Nick J Proudfoot. Ending the message: poly (A) signals then and now. *Genes & development*, 25(17):1770–1782, 2011.

[6] Bin Tian and James L Manley. Alternative cleavage and polyadenylation: the long and short of it. *Trends in Biochemical Sciences*, 38(6):312–320, 2013.

[7] Wei Wang, Zhi Wei, and Hongzhe Li. A change-point model for identifying 3' UTR switching by next-generation RNA sequencing. *Bioinformatics*, 30(15):2162–2170, 2014.

[8] Zheng Xia, Lawrence A Donehower, Thomas A Cooper, Joel R Neilson, David A Wheeler, Eric J Wagner, and Wei Li. Dynamic Analyses of Alternative Polyadenylation from RNA-Seq Reveal 3'-UTR Landscape Across 7 Tumor Types. *Nature Communications*, 5:5274, 2014.

User Manual

Download

APA-Scan is downloadable directly from github. Users need to have python (version 3.0 or higher) installed in their machine.

Required Softwares

(a) Python (v3.0 or higher)

(b) Samtools (v 0.1.8)* [This specific version is mandatory]

Required python packages

(a) Pandas

(b) Bio

(c) Scipy

(d) Numpy

(e) PeakUtils

Running APA-Scan

APA-Scan can handle both human and mouse data for detecting potential APA truncation sites. The tool is designed to follow the format of Refseq annotation and genome file from UCSC Genome Browser. Users need to have the following two files in the parent directory in order to run APA-Scan:
- Refseq annotation (.txt format)
- Genome fasta file (downloaded from UCSC genome browser)

APA-Scan comprises of two python scripts:
- APA-scan.py
- Make-plots.py

Run APA-scan.py

```
$ python3  APA-scan.py  annotation  ref_genome  input_dir1  input_dir2
-o  output_dir  -p  pas_dir1  pas_dir2
```

Example:

```
$ python3  APA-scan.py  annotation.txt  genome.fa  D://S1.bam  D://S2.bam
-o  Results -p  D://P1.bam  D://P2.bam
```

Options: (*denotes mandatory fields)

input1_dir*	Required field, directory of input1 RNA-seq data
input2_dir*	Required field, directory of input2 RNA-seq data
-o/-O	Denotes output directory. It is an optional field. If -o is not specified, the results will be generated inside of 'Output' folder.
-p/-P	P denotes whether the user gives the 3'-end-seq data or not. If -p is initialized, the next two fields after -p will be the directories of 3' end data for two samples. If -p is not specified, APA-Scan will automatically determine APA events according to its algorithm.

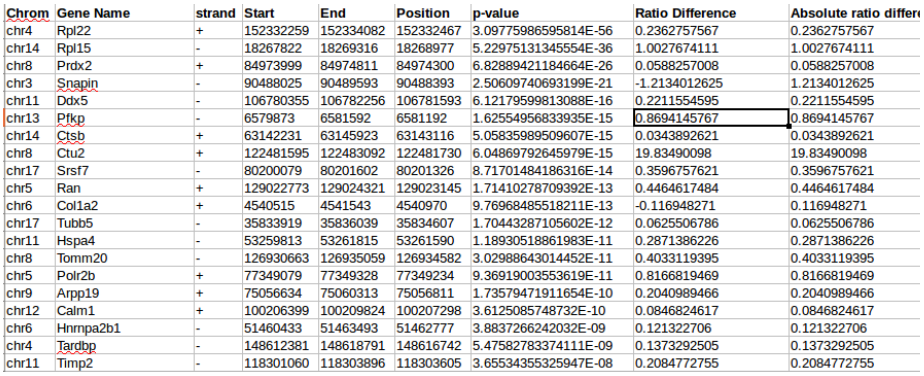
APA-Scan.py Results

APA-Scan will generate a spreadsheet in the output directory, with the following name:

- Result_PAS.csv [if the user provides the PAS data]
- Result.csv [if only RNA−seq input is provided], which contains the potential

transcript splice site for each region. APA-Scan will also generate some intermediary files in the output directory for reference purpose to the users.

The Result.csv [or Result_PAS.csv] file will contain the following fields (see image below) as long as all other information necessary to compute the association among two samples.



Chrom	Gene Name	strand	Start	End	Position	p-value	Ratio Difference	Absolute ratio differe
chr4	Rpl22	+	152332259	152334082	152332467	3.09775986595814E-56	0.2362757567	0.2362757567
chr14	Rpl15	-	18267822	18269316	18268977	5.22975131345554E-36	1.0027674111	1.0027674111
chr8	Prdx2	+	84973999	84974811	84974300	6.82889421184664E-26	0.0588257008	0.0588257008
chr3	Snapin	-	90488025	90489593	90488393	2.50609740693199E-21	-1.2134012625	1.2134012625
chr11	Ddx5	-	106780355	106782256	106781593	6.12179599813088E-16	0.2211554595	0.2211554595
chr13	Pfkp	-	6579873	6581592	6581192	1.62554956833935E-15	0.8694145767	0.8694145767
chr14	Ctsb	+	63142231	63145923	63143116	5.05835989509607E-15	0.0343892621	0.0343892621
chr8	Ctu2	+	122481595	122483092	122481730	6.04869792645979E-15	19.83490098	19.83490098
chr17	Srsf7	-	80200079	80201602	80201326	8.71701484186316E-14	0.3596757621	0.3596757621
chr5	Ran	+	129022773	129024321	129023145	1.71410278709392E-13	0.4464617484	0.4464617484
chr6	Col1a2	+	4540515	4541543	4540970	9.76968485518211E-13	-0.116948271	0.116948271
chr17	Tubb5	-	35833919	35836039	35834607	1.70443287105602E-12	0.0625506786	0.0625506786
chr11	Hspa4	-	53259813	53261815	53261590	1.18930518861983E-11	0.2871386226	0.2871386226
chr8	Tomm20	-	126930663	126935059	126934582	3.02988643014452E-11	0.4033119395	0.4033119395
chr5	Polr2b	+	77349079	77349328	77349234	9.36919003553619E-11	0.8166819469	0.8166819469
chr9	Arpp19	+	75056634	75060313	75056811	1.73579471911654E-10	0.2040989466	0.2040989466
chr12	Calm1	+	100206399	100209824	100207298	3.6125085748732E-10	0.0846824617	0.0846824617
chr6	Hnrnpa2b1	-	51460433	51463493	51462777	3.8837266242032E-09	0.121322706	0.121322706
chr4	Tardbp	-	148612381	148618791	148616742	5.47582783374111E-09	0.1373292505	0.1373292505
chr11	Timp2	-	118301060	118303896	118303605	3.65534355325947E-08	0.2084772755	0.2084772755

Run Make-plots.py

Command 1:

```
$ python3 Make-plots.py annotation ref_genome input_dir1 input_dir2
-o output_dir -p pas_dir1 pas_dir2
```

Example:

```
$ python3 Make-plots.py annotation.txt genome.fa  D://S1.bam
D://S2.bam -o Results -p D://P1.bam D://P2.bam
```

Command 2:
After executing the first command for a few seconds, **Make-plots.py** will ask the user to insert the region of interest in a specific format:
Chrom:GeneName:RegionStart-RegionEnd

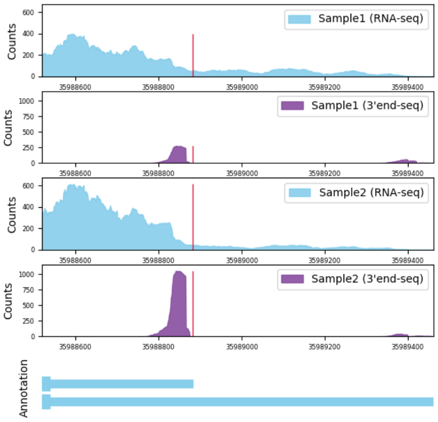
Parameters Explanation:
Chrom: Chromosome Name. Example: chr1
GeneName: denotes the gene ID or gene Name. Example: Tceb1
RegionStart: Start of the untranslated region
RegionEnd: End of the untranslated region

Example:

```
chr1:Tceb1:16641724-16643478
```

Make-plots.py Results

Make-Plots.py will generate a visual representation of the results shown in step 5, for each of the regions entered. The plot will illustrate the most significant transcript cleavage site with a red vertical bar on top of RNA-seq read data (and 3'end-seq if available). If the input parameters have 3'end-seq information along with the RNA-seq, then it will generate plots for both cases (See figure below). It will also show the UTR truncation point (annotated and unannotated) at the bottom panel.


Counts
Sample1 (RNA-seq)
Counts
Sample1 (3'end-seq)
Sample2 (RNA-seq)
Sample2 (3'end-seq)
Annotation