TAPAS: Tool for Alternative Polyadenylation Site Analysis (Supplementary Materials)

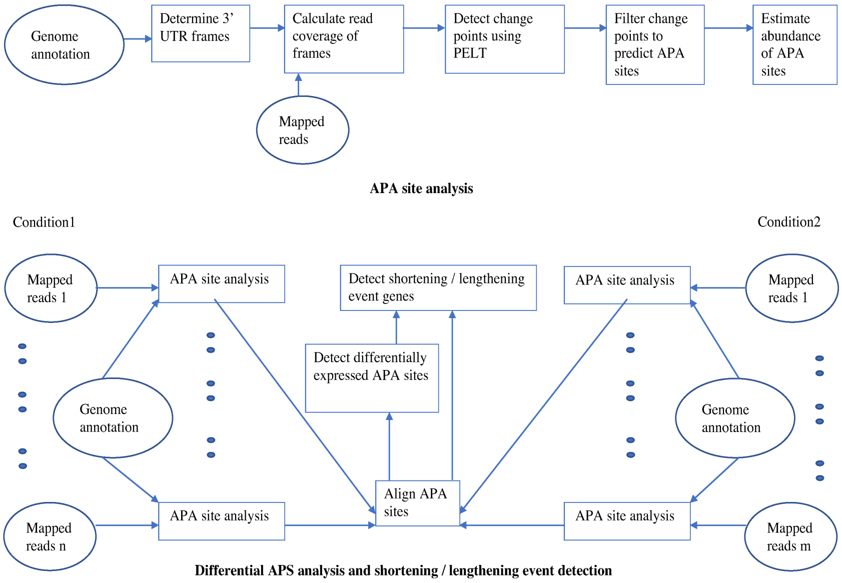
Ashraful Arefeen, Juntao Liu, Xinshu Xiao, and Tao Jiang

Table S1: Performance comparison in APA site detection on simulated data. The number of true APA sites is 21731.

Dataset (in million)	Tool name	Number of predicted APA sites	Correctly identified APA sites	Sensitivity (%)	Precision (%)
50	TAPAS	19453	16866	77.61	86.70
100	TAPAS	20712	18205	83.77	87.90
150	TAPAS	21335	18871	86.84	88.45
50	Cufflinks	25952	15117	69.56	58.25
100	Cufflinks	26032	16303	75.02	62.63
150	Cufflinks	25499	16779	77.21	65.80
50	IsoSCM	28152	11790	54.25	41.88
100	IsoSCM	29201	13583	62.51	46.52
150	IsoSCM	29600	14592	67.15	49.3
50	GETUTR	50818	15495	71.30	30.49
100	GETUTR	53226	16596	76.37	31.18
150	GETUTR	54577	17082	78.61	31.3

Table S2: Performance comparison in APA site detection on real data. Two flexible ranges (50 bps and 100 bps) are considered for matching a predicted APA site with a true one from 3′-Seq.

Number of true APA sites based on 3′-Seq	Tool name	Number of predicted APA sites	Correctly identified APA sites (50 bps)	Precision (%)	Correctly identified APA sites (100 bps)	Precision (%)
33751	TAPAS	33816	10429	30.84	12224	36.15
	Cufflinks	71502	5711	7.99	7956	11.13
	IsoSCM	36286	6354	17.51	7680	21.17
	GETUTR	62858	3111	4.95	6977	11.10



APA site analysis

Differential APS analysis and shortening / lengthening event detection

Figure S1: A flowchart of the TAPAS pipeline. In the differential expression analysis, we assume that n RNA-Seq replicates are given for each condition. In the figure, mapped reads also include read coverage information.

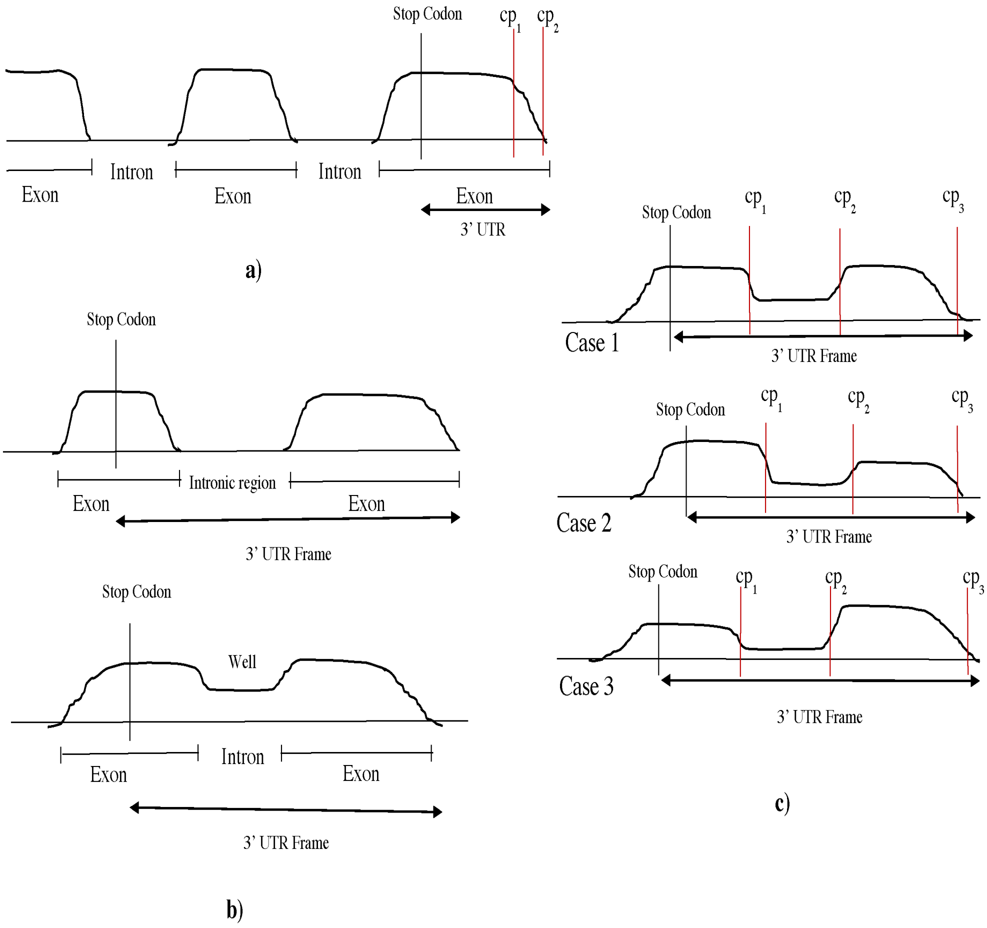


Figure S2: Some examples of filtration. (a) The PELT algorithm might output cp_1 as a change point even though the true APA site is cp_2, which is removed by TAPAS. (b) If a $3'$ UTR frame contains an intron (either annotated or novel), then a well might be created in the read coverage. (c) Three situations of the read coverage over the frame are illustrated. In case 1, the mean read coverages before and after the well are similar and TAPAS removes both change points cp_1 and cp_2 around the well. In case 2, the mean read coverage before the well is greater than the mean read coverage after the well and TAPAS keeps cp_1 as a potential APA site. In case 3, when the mean read coverage before the well is smaller than that after the well (which is not common), TAPAS would remove both change points as in the first case.

Table S3: Performance comparison in APA site detection on real data, when the prediction results of the tools compared are filtered by the 3′ UTR frames defined by TAPAS. Two flexible ranges (50 bps and 100 bps) are considered for matching a predicted APA site with a true one from 3′-Seq. The number of predicted APA sites of TAPAS is lowered to be closer to those of Cufflinks' and IsoSCM's. For a further comparison, Cufflinks is run with the reference transcriptome in RefSeq (*i.e.*, Cufflinks -g). Note that, given the number of APA sites predicted by Cufflinks -g, its performance should be directly compared with that of TAPAS provided in Table S2 rather than the numbers in this table.

Tool name	Number of predicted APA sites within frames	Correctly identified APA sites (50 bps flexible range)	Precision (%)	Correctly identified APA sites (100 bps flexible range)	Precision (%)
TAPAS	16313	8764	53.72	9764	59.85
Cufflinks	8719	3534	40.53	5034	57.74
Cufflinks -g	23594	9884	41.89	10838	45.94
IsoSCM	10016	4569	45.62	5606	55.97
GETUTR	23347	2289	9.80	5452	23.35

Table S4: Performance comparison in detecting internal APA sites located inside the 3′ UTR frames on real data.

Tool name	Correctly predicted internal APA sites (50 bps flexible range)	Sensitivity (%)	Correctly predicted internal APA sites (100 bps flexible range)	Sensitivity (%)
TAPAS	7598	46.69	8302	51.01
Cufflinks	3906	24.00	5520	33.92
IsoSCM	4640	28.51	5586	34.32
GETUTR	2512	15.43	5579	34.28

Table S5: Performance comparison in APA site detection on real data. Two flexible ranges (50 bps and 100 bps) are considered for matching a predicted APA site with a true one from PAS-Seq.

Number of true APA sites based on PAS-Seq	Tool name	Number of predicted APA sites	Correctly identified APA sites (50 bps)	Precision (%)	Correctly identified APA sites (100 bps)	Precision (%)
50148	TAPAS	33816	26336	77.88	29346	86.78
	Cufflinks	71502	12338	17.26	17290	24.18
	IsoSCM	36286	17606	47.38	19919	54.89
	GETUTR	62858	6253	9.95	15442	24.57

Table S6: Performance comparison in the detection of genes with differentially expressed (DE) APA sites on simulated data. The number of genes with actual DE APA sites is 1254, and each such gene contains only one DE APA sites. Since DEXSeq is designed for differential splicing (DS) rather than DE analysis [Liu *et al.*, 2014, Soneson *et al.*, 2016], we consider DE genes with at least two transcripts (298 in total) as the benchmark when evaluating the performance of DEXSeq. Here, Cuffdiff_anno = Cuffdiff with annotation.

Dataset (in million)	Tool name	Number of detected genes with DE APA sites	Correctly identified genes with DE APA sites	Sensitivity (%)	Precision (%)
30	TAPAS	1282	955	76.16	74.49
50	TAPAS	1329	1048	83.57	78.86
100	TAPAS	1308	1119	89.23	85.55
150	TAPAS	1317	1139	90.83	86.48
30	Cuffdiff	1377	999	79.67	72.55
50	Cuffdiff	1388	1011	80.62	72.84
100	Cuffdiff	1429	1017	81.10	81.10
150	Cuffdiff	1446	1012	80.70	69.99
30	Cuffdiff_anno	1158	1022	81.50	88.26
50	Cuffdiff_anno	1180	1046	83.41	88.64
100	Cuffdiff_anno	1188	1057	84.29	88.97
150	Cuffdiff_anno	1200	1063	84.47	88.58
30	DESeq	1202	1129	90.03	93.93
50	DESeq	1210	1144	91.23	94.55
100	DESeq	1197	1124	89.63	93.90
150	DESeq	1235	1141	90.99	92.39
30	DEXSeq	281	198	66.44	70.46
50	DEXSeq	278	211	70.81	75.90
100	DEXSeq	268	215	72.15	80.22
150	DEXSeq	273	216	72.48	79.12

Table S7: Performance comparison in the detection of genes with shortening/lengthening events on simulated data. The actual number of genes with shortening/lengthening events is 674.

Dataset (in million)	Tool name	Number of predicted event genes	Correctly determined event genes	Sensitivity (%)	Precision (%)
50	TAPAS	598	444	65.88	74.25
100	TAPAS	632	502	74.48	79.43
150	TAPAS	631	506	75.07	80.19
50	DaPars	727	422	62.61	58.05
100	DaPars	645	426	63.20	66.05
150	DaPars	618	443	65.73	71.68
50	ChangePoint	421	125	18.55	29.69
100	ChangePoint	525	125	18.55	23.81
150	ChangePoint	509	138	20.47	27.11

Table S8: Performance comparison in the detection of genes with shortening/lengthening events on real data.

Tool name	Shortening /lengthening event gene identified by tool	Precision (%)
TAPAS	872	61.7
DaPars	808	39.85
ChangePoint	734	34.33

Table S9: Comparison of time (in minutes) and peak memory (in gigabytes) usage among the APA site detection tools on the simulated dataset with 50 million reads used in Section 3.1. Here, the running time of TAPAS includes the calculation of read coverage by SAMtools.

Tool name	Time (min)	Memory (GB)
TAPAS	121	16.62
Cufflinks	97	1.00
IsoSCM	103	3.67
GETUTR	106	19.78

Table S10: Comparison of time and peak memory usage among the tools for shortening/lengthening analysis on the simulated dataset with 50 millions reads used in Section 3.3. Again, the running time of TAPAS includes the calculation of read coverage by SAMtools.

Tool name	Time (min)	Memory (GB)
TAPAS	803	7.70
TAPAS (parallel)	81	7.70
DaPars	49	3.99
ChangePoint	1876	19.55

Algorithm 1 The PELT method for finding change points in a $3'$ UTR frame.

```
procedure PELTMethod(y, C, $\gamma$)
    Input:
    $y \rightarrow$ read coverage of a $3'$ UTR frame, $(y_1, y_2, \ldots, y_n)$
    $C \rightarrow$ twice negative log-likelyhood cost function on $y$
    $\gamma \rightarrow$ penalty

    Initialize:
    $F(0) = -\gamma$
    $cp(0)$ =NULL
    $R_1 = \{0\}$

    for $t^* = 1, \ldots, n$ do
        $F(t^*) = min_{t \in R_{t^*}}[F(t) + C(y_{t+1:t^*}) + \gamma]$
        $t^1 = arg\{min_{t \in R_{t^*}}[F(t) + C(y_{t+1:t^*}) + \gamma]\}$
        $cp(t^*) = [cp(t^1), t^1] - \{0\}$
        $R_{t^*+1} = \{t^*, \{t \in R_{t^*} : F(t) + C(y_{t+1:t^*}) < F(t^*)\}\}$
    $cp(n) = [cp(n), n]$
    Output: change points, $cp(n)$
```

Algorithm 2 Filtration of change points found by PELT when the read coverage of a $3'$ UTR frame increases (or decreases) gradually.

```
procedure FilterRedundantChangePoints(cp, coverage, strand)
    Input:
    $cp \rightarrow$ change points of a $3'$ UTR frame
    $coverage \rightarrow$ read coverage of the $3'$ UTR frame
    $strand \rightarrow$ strand of the $3'$ UTR frame
    if strand = positive then
        for each pair of consecutive change points, $(cp_{i-1}, cp_i)$ do
            if most of the base positions between $cp_{i-1}$ and $cp_i$ have decreasing coverage then
                remove $cp_{i-1}$ from the list of APA sites
    else
        for each pair of consecutive change points, $(cp_i, cp_{i+1})$ do
            if most of the base positions between $cp_i$ and $cp_{i+1}$ have increasing coverage then
                remove $cp_{i+1}$ from the list of APA sites
```

Algorithm 3 Detection and removal of change points around a well.

```
procedure FilterChangePointsAroundWell(cp, coverage, strand)
    Input:
    $cp \rightarrow$ change points of a $3'$ UTR frame. These change points divide the frame into segments
    $coverage \rightarrow$ read coverage of the $3'$ UTR frame
    $strand \rightarrow$ strand of the $3'$ UTR frame

    $M \leftarrow$ mean coverage of segments
    if strand = positive then
        for each mean $m_i$ in $M$ do
            if $m_{i-1} > m_i < m_{i+1}$ then
                if $m_{i-1} = m_{i+1}$ then
                    remove change points between $m_{i-1}$, $m_i$ and $m_i$, $m_{i+1}$
                else if $m_{i-1} > m_{i+1}$ then
                    remove change points between $m_i$ and $m_{i+1}$
                else
                    remove change point between $m_{i-1}$, $m_i$ and $m_i$, $m_{i+1}$
    else
        for each mean $m_i$ in $M$ do
            if $m_{i-1} > m_i < m_{i+1}$ then
                if $m_{i-1} = m_{i+1}$ then
                    remove change points between $m_{i-1}$, $m_i$ and $m_i$, $m_{i+1}$
                else if $m_{i-1} < m_{i+1}$ then
                    remove change points between $m_{i-1}$ and $m_i$
                else
                    remove change point between $m_{i-1}$, $m_i$ and $m_i$, $m_{i+1}$
```

Algorithm 4 EM algorithm for estimating the abundance of alternative 3′ UTRs.

```
procedure AbundanceCalculator($T$, $l$, $R$)
    Input:
    $T \rightarrow$ set of all possible alternative 3′ UTRs in a frame
    $l \rightarrow$ set of lengths of those alternative 3′ UTRs
    $R \rightarrow$ set of reads mapped in the 3′ UTR frame

    Assign random values to all $\rho_t$, where $t \in T$ and $\rho_t$ is the abundance of $t$
    while not converged do
        initialize all $read_t$ to 0, where $read_t$ is the read count for $t$
        for each read $r$ in $R$ do
            $T_r \rightarrow$ set of alternative 3′ UTRs containing read $r$
            for each alternative 3′ UTR $t$ ($t \in T_r$) do
                $read_t = read_t + \frac{\rho_t}{\sum_{u \in T_r} \rho_u}$
        $s = \sum_{t \in T} \frac{read_t}{(l_t - l_r + 1)}$
        for each alternative 3′ UTR $t$ do
            $\rho_t = \frac{read_t}{(l_j - l_r + 1) \times s}$
    $RC \leftarrow$ calculate the abundance (read counts) of all the 3′ UTRs (of the given frame) from $\rho$
    Output: $RC$
```

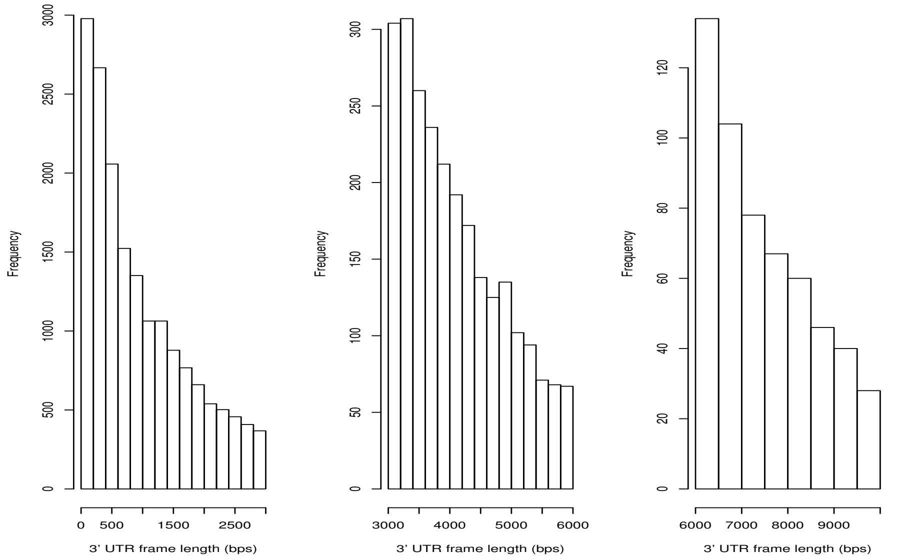


Figure S3: Length distribution of the 3′ UTR frames extract from the human RefSeq annotation GRCh37. The 3′ UTR frames have lengths ranging from 2 bps to 238,767 bps, with the average being $1,770.786$ bps.

Table S11: Performance comparison between TAPAS and 3P-Seq in APA site detection on mouse liver data. Paired-end RNA-Seq reads from standard polyA+ libraries for mouse liver (SRX196268) were downloaded from NCBI and mapped by TopHat2 to the mouse genome. For performance evaluation, a 3′-Seq dataset for mouse liver (GSM747483) was also downloaded from NCBI and used as benchmark. We ran TAPAS on the mapped reads and compared its predicted APA sites against the benchmark. As a comparison, we downloaded the 3P-Seq data for mouse liver (GSM1268948) from NCBI. Among the 29932 APA sites reported in the 3-Seq data, TAPAS and 3P-Seq identified 10900 and 19480 sites, respectively. In terms of sensitivity, 3P-Seq outperforms TAPAS; but TAPAS outperforms 3P-Seq in terms of precision. Note that TAPAS uses standard RNA-Seq data which is very popular and easy to perform while 3P-Seq requires complex biological steps and large amounts of RNA for its analysis [Kim *et al.*, 2015].

Number of APA sites in 3′-Seq data	Tool name	Number of output APA sites	Overlap with 3′-Seq (100 bps flexible range)	Sensitivity (%)	Precision (%)
29932	TAPAS	25147	10900	36.42	43.35
	3P-Seq	82551	19480	65.08	23.60

Table S12: Versions of the other tools compared in the experiments.

Tool name	Version
IsoSCM	2.0.11
GETUTR	1.0.2
Cufflinks	2.2.1
Cuffdiff	2.2.1
DESeq	1.9.12
DEXSeq	0.1.25

Commands for Running the Tools Compared in the Experiments

TAPAS

```
./APA_sites_detection -ref ANNOTATION_FILE -cov READ_COVERAGE_FILE -l READ_LENGTH -o OUTPUT_FILE
```

As also explained on its Github page[1], the expected input of TAPAS consists of a genomic or transcriptomic annotation file (from UCSC), a read coverage file (generated using SAMtools) and the read length, and its output includes a list of predicted APA sites in all extracted 3′ UTR frames.

IsoSCM

```
java -Xmx102400m -jar IsoSCM-2.0.11.jar assemble -coverage false -bam BAM_FILE
-base OUTPUT_FILE_NAME -s unstranded -min_terminal 50 -min_fold 0.08 -jnct_alpha 0.05
```

GETUTR

```
python GETUTR.1.0.2/GETUTR.py -i BAM_FILE_NAME -o OUTPUT_FILE_NAME -m 10 -r ANNOTATION_FILE
```

Cufflinks

```
cufflinks -p 1 -o OUTPUT_FILE --overlap-radius 75 BAM_FILE
```

[1] https://github.com/arefeen/TAPAS

Cuffdiff

```
cuffdiff -o OUTPUT_FILE -p 2 -FDR 0.1 -L C1,C2 -m 76 -s 1 -max-bundle-frags 20000000
ANNOTATION_FILE SET_OF_CONDITION_ONE_FILES SET_OF_CONDITION_TWO_FILES
```

DESeq and DEXSeq

These tools are run with their default settings in Bioconductor.

Dapars

```
python DaPars_Extract_Anno.py -b INPUT_BED_FILE -s ANNOTATION_FILE -o OUTPUT_BED_FILE
python DaPars_main.py CONFIGURATION_FILE
```

ChangePoint

```
perl ChangePoint/change_point.pl -c CONDITION_ONE_BAM -t CONDITION_TWO_BAM -r 2 -n 5 -a 0.1 -x
51200m -g ANNOTATION_FILE -d s -o OUTPUT_FILE
perl ChangePoint/change_point.pl -c CONDITION_ONE_BAM -t CONDITION_TWO_BAM -r 2 -n 5 -a 0.1 -x
51200m -g ANNOTATION_FILE -d l -o OUTPUT_FILE
```

References

[Kim *et al.*, 2015] Kim, M. *et al.* (2015) Global estimation of the 3′ untranslated region landscape using RNA sequencing. *Methods*, **83**, 111-117.

[Liu *et al.*, 2014] Liu, R. *et al.* (2014) Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics*, **15**(1), 364.

[Soneson *et al.*, 2016] Soneson, C. *et al.* (2016) Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, **17**, 12.