

APA-Scan User Manual

Naima Ahmed Fahmi, Jae-Woong Chang, Heba Nasserreddeen, Khandakar Tanvir Ahmed,
Deliang Fan, Jeongsik Yong and Wei Zhang

1. Download

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

2. Required Softwares

- a. [Python](#) (v3.0 or higher)
- b. [Samtools \(v 0.1.8\)](#)* [This specific version is mandatory]

3. Required python packages

- a. [Pandas](#)
- b. [Bio](#)
- c. [Scipy](#)
- d. [Numpy](#)
- e. [Peakutils](#)

4. 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)

-s/-S*	Species name. APA-Scan can handle human and mouse in the current version. Users have to specify h for human and m for mouse.
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. Its 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 determines APA events according to its algorithm.

5. APA-Scan.py Results

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

- Result_PAS.csv [if user provides the PAS data]
 - Result.csv [if only RNA-seq input is provided]
- , which will contain 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 differ
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	Snapi	-	90488025	90489593	90488393	2.50609740693199E-21	-1.2134012625	1.2134012625
chr11	Ddx5	-	106780355	106782256	106781593	6.12179599813088E-16	0.2211554595	0.2211554595
chr13	Pfkfb	-	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	Hnmpa2b1	-	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

6. 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: Chrom:GeneName:RegionStart-RegionEnd

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`

7. Make-plots.py Results

Make-Plots.py will generate a visual representation of the results shown in step 5, for each of the region 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.

