APA-Scan User Manual

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

1. About

APA-Scan is a computational tool which can detect and visualize genome-wide 3'-UTR APA events. APA-Scan integrates both 3'-end-seq (an RNA-seq method with a specific enrichment of 3'-ends ofmRNA) data and the location information of predicted canonical PASs with RNA-seq data to improve the quantitative definition of genome-wide UTR-APA events. It is also advantageous in producing high quality plots of the user defined events.

2. Download

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

3. Required Softwares

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

4. Required python packages (Can install using pip, or other process)

a. Pandas: $ pip install pandas
b. Bio: $ pip install biopython
c. Scipy: $ pip install scipy
d. Numpy: $ pip install numpy
e. Peakutils: pip install PeakUtils

5. Run 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)

RefSeq annotation can be downloaded from UCSC Genome browser using the following setup in *Tools* -> *Table browser*:

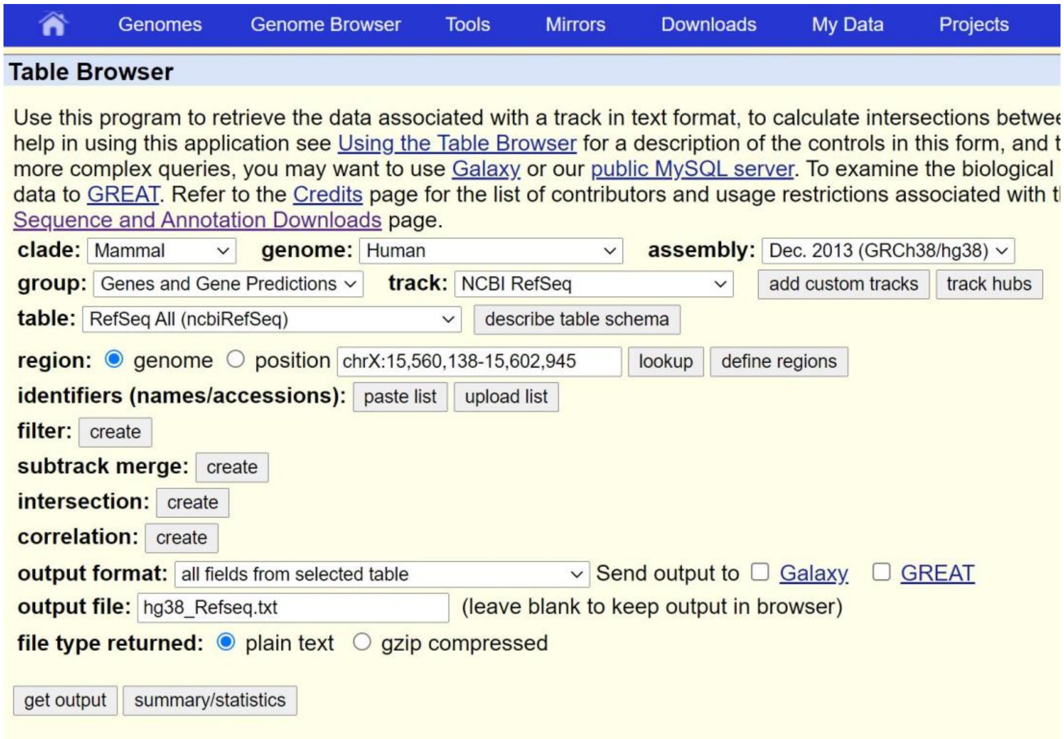


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections betwee
help in using this application see Using the Table Browser for a description of the controls in this form, and t
more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological
data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with t
Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions **track:** NCBI RefSeq add custom tracks track hubs

table: RefSeq All (ncbiRefSeq) describe table schema

region: ● genome ○ position chrX:15,560,138-15,602,945 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

subtrack merge: create

intersection: create

correlation: create

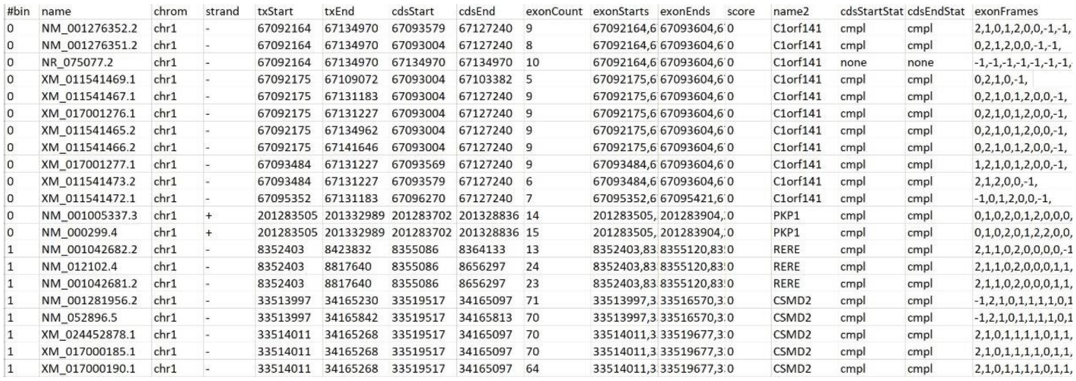
output format: all fields from selected table Send output to □ Galaxy □ GREAT

output file: hg38_Refseq.txt (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

The annotation.txt file downloaded from the UCSC Genome browser will have the following columns:



#bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	score	name2	cdsStartStat	cdsEndStat	exonFrames
0	NM_001276352.2	chr1	-	67092164	67134970	67093579	67127240	9	67092164,6	67093604,6	0	C1orf141	cmpl	cmpl	2,1,0,1,2,0,0,-1,-1,
0	NM_001276351.2	chr1	-	67092164	67134970	67093004	67127240	8	67092164,6	67093604,6	0	C1orf141	cmpl	cmpl	0,2,1,2,0,0,-1,-1,
0	NR_075077.2	chr1	-	67092164	67134970	67134970	67134970	10	67092164,6	67093604,6	0	C1orf141	none	none	-1,-1,-1,-1,-1,-1,-1,
0	XM_011541469.1	chr1	-	67092175	67109072	67093004	67103382	5	67092175,6	67093604,6	0	C1orf141	cmpl	cmpl	0,2,1,0,-1,
0	XM_011541467.1	chr1	-	67092175	67131183	67093004	67127240	9	67092175,6	67093604,6	0	C1orf141	cmpl	cmpl	0,2,1,0,1,2,0,0,-1,
0	XM_017001276.1	chr1	-	67092175	67131227	67093004	67127240	9	67092175,6	67093604,6	0	C1orf141	cmpl	cmpl	0,2,1,0,1,2,0,0,-1,
0	XM_011541465.2	chr1	-	67092175	67134962	67093004	67127240	9	67092175,6	67093604,6	0	C1orf141	cmpl	cmpl	0,2,1,0,1,2,0,0,-1,
0	XM_011541466.2	chr1	-	67092175	67141646	67093004	67127240	9	67092175,6	67093604,6	0	C1orf141	cmpl	cmpl	0,2,1,0,1,2,0,0,-1,
0	XM_017001277.1	chr1	-	67093484	67131227	67093569	67127240	9	67093484,6	67093604,6	0	C1orf141	cmpl	cmpl	1,2,1,0,1,2,0,0,-1,
0	XM_011541473.2	chr1	-	67093484	67131227	67093579	67127240	6	67093484,6	67093604,6	0	C1orf141	cmpl	cmpl	2,1,2,0,0,-1,
0	XM_011541472.1	chr1	-	67095352	67131183	67096270	67127240	7	67095352,6	67095421,6	0	C1orf141	cmpl	cmpl	-1,0,1,2,0,0,-1,
0	NM_001005337.3	chr1	+	201283505	201332989	201283702	201328836	14	201283505,	201283904,	0	PKP1	cmpl	cmpl	0,1,0,2,0,1,2,0,0,0,
0	NM_000299.4	chr1	+	201283505	201332989	201283702	201328836	15	201283505,	201283904,	0	PKP1	cmpl	cmpl	0,1,0,2,0,1,2,2,0,0,
1	NM_001042682.2	chr1	-	8352403	8423832	8355086	8364133	13	8352403,83	8355120,83	0	RERE	cmpl	cmpl	2,1,1,0,2,0,0,0,0,-1
1	NM_012102.4	chr1	-	8352403	8817640	8355086	8656297	24	8352403,83	8355120,83	0	RERE	cmpl	cmpl	2,1,1,0,2,0,0,0,1,1,
1	NM_001042681.2	chr1	-	8352403	8817640	8355086	8656297	23	8352403,83	8355120,83	0	RERE	cmpl	cmpl	2,1,1,0,2,0,0,0,1,1,
1	NM_001281956.2	chr1	-	33513997	34165230	33519517	34165097	71	33513997,3	33516570,3	0	CSMD2	cmpl	cmpl	-1,2,1,0,1,1,1,1,0,1
1	NM_052896.5	chr1	-	33513997	34165842	33519517	34165813	70	33513997,3	33516570,3	0	CSMD2	cmpl	cmpl	-1,2,1,0,1,1,1,1,0,1
1	XM_024452878.1	chr1	-	33514011	34165268	33519517	34165097	70	33514011,3	33519677,3	0	CSMD2	cmpl	cmpl	2,1,0,1,1,1,1,0,1,1,
1	XM_017000185.1	chr1	-	33514011	34165268	33519517	34165097	70	33514011,3	33519677,3	0	CSMD2	cmpl	cmpl	2,1,0,1,1,1,1,0,1,1,
1	XM_017000190.1	chr1	-	33514011	34165268	33519517	34165097	64	33514011,3	33519677,3	0	CSMD2	cmpl	cmpl	2,1,0,1,1,1,1,0,1,1,

APA-Scan has two python scripts: APA-Scan.py, Make-Plots.py
And 1 configuration file: configuration.ini

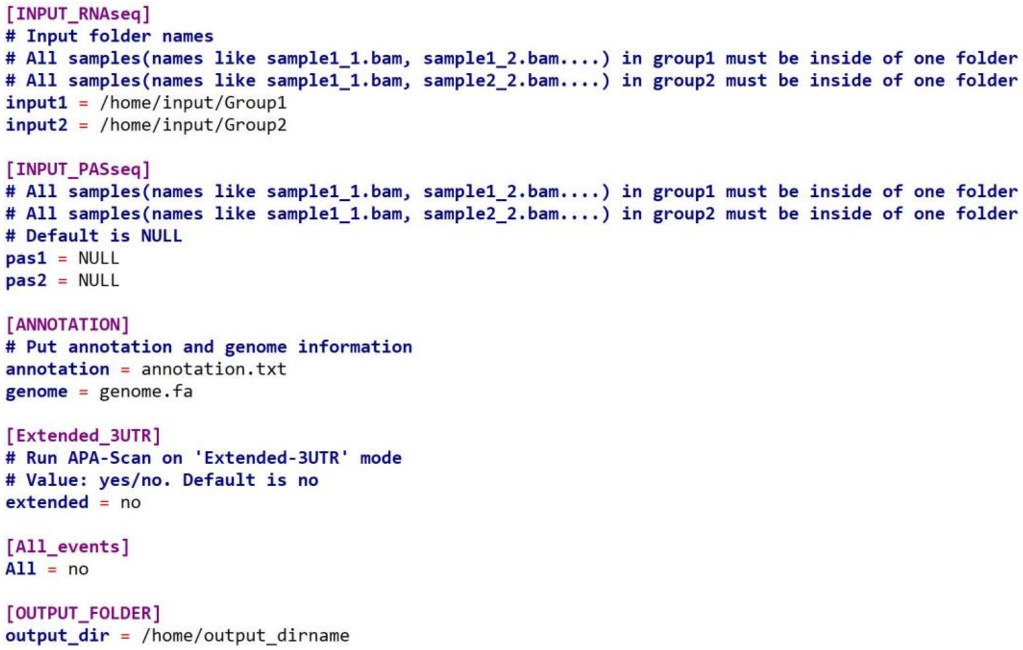
The configuration file allows the users to specify:
1) the directories of the input samples,
2) the species to be analyzed, and
3) the directory of the folder where all output files will be stored.

APA-Scan supports the analysis of multiple samples that belong to two different groups- all BAM files inside the input1 directory will be considered as part of the first group, and all BAM files inside the input2 directory will be considered as part of the second group. It is required to have at least one BAM file in each input directory.

Running Parameters in the configuration.ini file: (* refers to a mandatory field)

species*:	Species name (human/mouse)
input1* :	Directory containing the first group of samples with RNA-seq data.
input2* :	Directory containing the second group of samples with RNA-seq data.
pas1* :	Directory containing the first group of samples with 3'-end-seq data. Default is NULL
pas2* :	Directory containing the second group of samples with 3'-end-seq data. Default is NULL
extended* :	APA-Scan will run on `Extended 3UTR' mode and it will search for APA sites upto 10kb downstream of the annotated transcript. Value: yes or no
All* :	If selected 'yes', APA-Scan will report all the candidate cleavage sites of a gene, whether they are significant or not. Otherwise, APA-Scan will report the most significant event for each gene [default]. Value: yes or no
annotation* :	RefSeq annotation file, downloaded from UCSC Genome Browser, in .txt format
genome* :	Genome fasta file, in .fa format
output_dir* :	Output directory. Users can specify the desired output directory for writing the results. [Optional]

An example of the congiration.ini file is provided below:

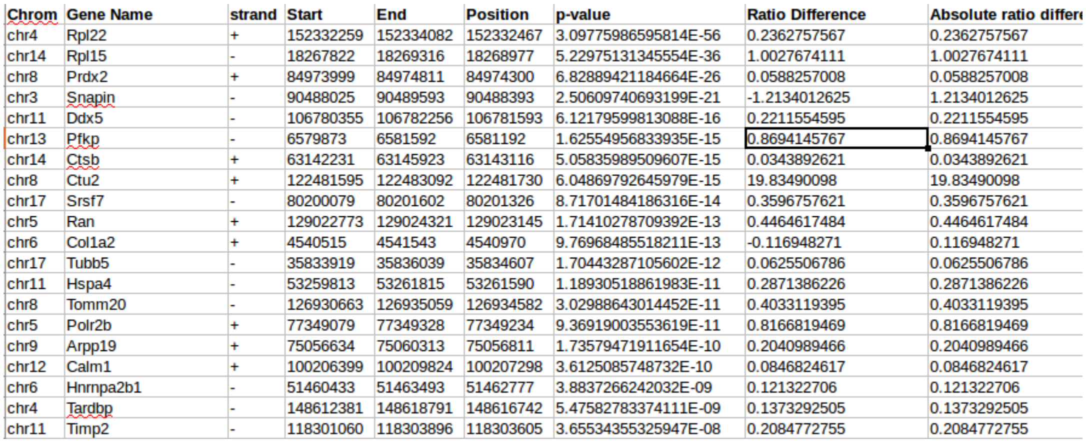


```
[INPUT_RNAseq]
# Input folder names
# All samples(names like sample1_1.bam, sample1_2.bam....) in group1 must be inside of one folder
# All samples(names like sample1_1.bam, sample2_2.bam....) in group2 must be inside of one folder
input1 = /home/input/Group1
input2 = /home/input/Group2

[INPUT_PASseq]
# All samples(names like sample1_1.bam, sample1_2.bam....) in group1 must be inside of one folder
# All samples(names like sample1_1.bam, sample2_2.bam....) in group2 must be inside of one folder
# Default is NULL
pas1 = NULL
pas2 = NULL

[ANNOTATION]
# Put annotation and genome information
annotation = annotation.txt
genome = genome.fa

[Extended_3UTR]
# Run APA-Scan on 'Extended-3UTR' mode
# Value: yes/no. Default is no
extended = no

[All_events]
All = no

[OUTPUT_FOLDER]
output_dir = /home/output_dirname
```

Once the parameters have been specified in the configuration file, the user will enter the following command to run APA-Scan:

```
$ python3   APA-Scan.py
```

APA-Scan.py will generate several intermediary files in the output directory. After computing the significance of the association between the two groups of samples, the final results will be written in the file named Group1_Vs_Group2.csv. The following image shows some of the generated fields in Group1_Vs_Group2.csv:



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

The column ‘p-value’ defines the significance of the UTR-APA events. In the ‘Ratio difference’ column, a large positive ratio difference indicates a potential UTR truncation occurred in condition 2, whereas a negative ratio difference with a large absolute value indicates a potential UTR-APA event in condition 1.

6. **Run Make-plots.py**

Make-plots.py also requires the same configuration file to run. It will use the input and output directories listed in the configuration file and prepare a read coverage plot along with the 3'-UTR annotation based on user defined region.

```
python3  Make-plots.py
```

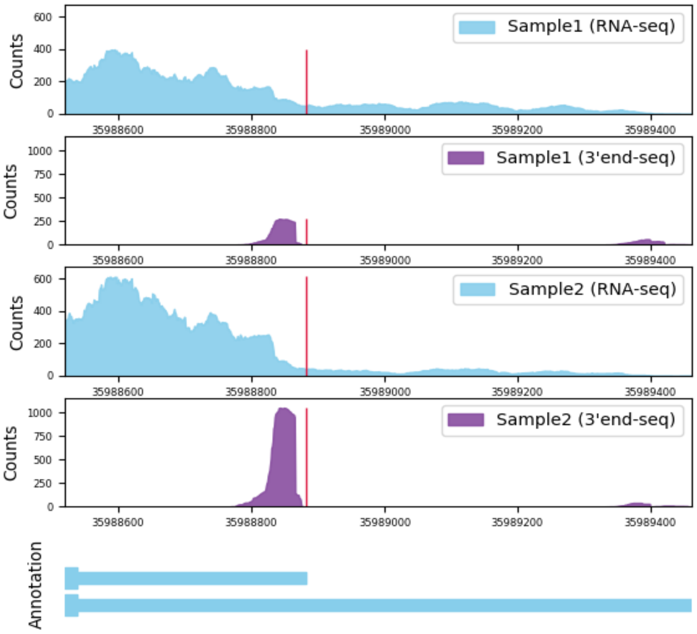
After executing this command above for a few seconds, **Make-plots.py** will ask the user to insert the region of interest in a specific format:
Chrom:GeneName:RegionStart-RegionEnd

Chrom :	Name of the chromosome
GeneName :	Name of the gene
RegionStart :	Start position of the region
Region End :	End position of the region

Example:

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

Make-Plots.py will generate a visual representation of the results shown 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
600
400
200
0
Sample1 (RNA-seq)
35988600
35988800
35989000
35989200
35989400
Counts
1000
750
500
250
0
Sample1 (3'end-seq)
35988600
35988800
35989000
35989200
35989400
Counts
600
400
200
0
Sample2 (RNA-seq)
35988600
35988800
35989000
35989200
35989400
Counts
1000
750
500
250
0
Sample2 (3'end-seq)
35988600
35988800
35989000
35989200
35989400
Annotation