

# README for MADS: Microarray Analysis of Differential Splicing version 1.0

<http://biogibbs.stanford.edu/~yxing/MADS/>

---

## Table of Contents

---

1. Overview
2. Usage
3. Input Files and Parameters
4. Output Files
5. Contact

---

## 1. Overview

---

MADS is a tool to discover differential alternative splicing events from exon tiling microarray data. The principle of MADS is to increase the precision of exon-level and gene-level expression estimates by correcting, as much as possible, noise in observed probe intensities due to background and cross-hybridization. Our software incorporates a series of novel algorithms motivated by the “probe-rich” design of exon-tiling arrays, such as background correction, iterative probe selection and removal of sequence-specific cross-hybridization to off-target transcripts.

We used MADS to analyze Affymetrix Exon 1.0 array data on a mouse neuroblastoma cell line after shRNA-mediated knockdown of the splicing factor PTB ([GSE11344](#)). The results of this analysis are published at RNA, 2008, 14(8): 1470-1479.

---

## 2. Usage

---

Python and R are required for running MADS. Python package [rpy](#) (R from Python) is also required.

For a test run of MADS, download the [sample data](#) and un-zip the tar-zipped file into a folder.

Run the MADS.py with the following command:

```
python MADS.py PTB_expression.xls sample_info probe_intensity/ mm8Probeset
Gene_Info_Mouse.xls Mouse_CrossHybOutput.txt output.txt -h 0.55 -x -0.9 -f 2.0 -g
500 -t 0
```

A description of the input and output files is given in the next section.

---

### 3. Input Files and Parameters

---

Here is a brief description of the input files and parameters:

```
python MADS.py Expression_index_file Sample_info Probe_intensity_directory
Probeset_Annotation Gene_info Cross_hybrid_file Output_file [-h probe cross-
hybridization cutoff] [-x extreme value cutoff] [-f fold change cutoff] [-g gene expression
cutoff] [-t unpaired or paired t test]
```

Expression\_index\_file: gene expression index file

Sample\_info: grouping information of sample files

Probe\_intensity\_directory: directory with background corrected intensities of individual probes

Probeset\_Annotation: list of probesets on the array and their annotations

Gene\_info: gene symbols and names

Cross\_hybrid\_file: list of probes that cross-hybridize to off-target transcripts

Output\_file: output file of analysis result

[-h probe cross-hybridization cutoff]: cutoff of Pearson correlation co-efficient for filtering of cross-hybridizing probes. We define a probe to be cross-hybridizing if there is an off-target transcript within 3bp mismatches, and if the computed Pearson correlation coefficient is above a user-defined cutoff (the default value is 0.55). Cross-hybridizing probes are regarded as unreliable and filtered from further exon-level analysis.

[-x extreme value cutoff]: cutoff for filtering of extreme-value probes. We detect probes whose intensities are higher than a user-defined cutoff percentile (the default value is 90%) of all other core probes of the gene in at least one of the two sample groups. Such extreme-value probes are filtered from further exon-level analysis.

[-f fold change cutoff]: cutoff of fold change in gene expression levels. Genes whose fold changes between two sample groups are higher than the user-defined cutoff (the default value is 2.0) are removed from further analysis.

[-g gene expression cutoff]: cutoff of gene expression index. Lowly expressed genes whose expression indices are below a given user-defined cutoff (the default value is 500) are removed from further analysis.

[-t unpaired or paired t test]: option for doing unpaired t test or paired t test (0=unpaired, 1=paired). When the experiment design has paired samples, paired t test will give more accurate estimation.

The first 7 input files/parameters are mandatory for MADS program while the last 4 parameters are optional.

Before running MADS to detect differentially spliced exons, the user first need to use ProbeEffects and ProbeSelect scripts to perform background-correction, normalization and expression index computation of Exon array data. These scripts and user instructions can be downloaded from [GeneBASE](#). The following input files are the output files of ProbeEffects and ProbeSelect scripts: Expression\_index\_file, Probe\_intensity\_directory.

The details about the input files/parameters are as follow:

**[Expression\_index\_file]** is gene expression index file. It is obtained by performing probe selection and expression index calculation using the [ProbeSelect](#) script. An example file is PTB\_expression.xls. It looks like:

TranscriptCluster	H1.1	H1.2	H1.3	siPTB.1	siPTB.2	siPTB.3
6747308	925	1015	1301	1287	1358	1406
6747309	2562	3559	2906	2907	2645	2901
6747314	2119	2340	2744	3030	2932	2751
6747326	11	7	-1	15	21	45
6747343	581	652	653	682	727	576
6747354	56	73	99	422	521	406

The first column is the transcript cluster id. The next columns are the transcript cluster expression index of every sample.

For reliable estimates of gene expression levels, we recommend running ProbeSelect on a diverse set of samples covering different tissues or cell types. For example, in the analysis of the PTB-knockdown dataset [RNA (doi:10.1261/rna.1070208)], we perform probe selection after combining the PTB dataset with [Affymetrix public Exon array data](#) for 11 mouse tissues. Then we write the expression indices estimated for the six PTB samples into a separate file (PTB\_expression.xls) for downstream exon-level analysis.

**[SAMPLE\_INFO]** contains the grouping information of sample files. It must be created by the user. Samples with the same names are considered in the same group. The order of names in this file **MUST** be consistent with the order of sample names in [Expression\_index\_file]. An example file is sample\_info. It looks like:

h1 h1 h1 ptb ptb ptb
----------------------

The first 3 columns indicate that the first 3 samples are in the same group, while the last 3 columns indicate that the last 3 samples are in another group. **NOTE:** For the paired test, which is  $-t = 1$  in the option, an additional line should be added into the sample\_info file to indicate how samples are paired. Samples with the same numeric identifiers in this line are paired samples. Here is an example for sample\_info in paired t test:

h1 h1 h1 ptb ptb ptb
1 2 3 1 2 3

**[PROBE\_INTENSITY\_DIRECTORY]** is the directory which contains the background corrected intensities of individual probes. It is obtained by performing background correction and expression index calculation using the [ProbeEffects and ProbeSelect](#) scripts. An example file is probe\_intensity.tar.gz It needed to be un-zipped into folder probe\_intensity/. An example file of one transcript cluster looks like:

6757994	H1.1	H1.2	H1.3	siPTB.1	siPTB.2	siPTB.3
1593144 4311468 core	229.093	247.392	184.212	131.547	259.233	187.226
2295002 4311468 core	110.682	192.225	162.718	12.6206	158.158	246.777
5984044 4311468 core	94.7205	106.729	209.887	215.907	87.1736	107.079
1477163 4338597 core	783.14	1051.93	761.808	850.619	745.819	948.291
2801311 4338597 core	188.514	176.074	164.379	97.6128	221.865	161.414
4138581 4338597 core	1959.24	1588.52	2016.17	1466.16	1636.97	1779.47

The first column is the probe\_id|probeset\_id|type of the transcript cluster (6757994). The next columns are the probe intensities of this probe in different samples. The order of samples in these files needs to be consistent with the order of samples in Expression\_index\_file.

**[PROBESET\_ANNOTATION]** is the list of probesets on the array and their annotations (e.g. genomic coordinates). Example files are hg18Probeset for human and mm8Probeset for mouse. It looks like:

probeset_id	transcript_cluster_id	chromosome	strand	start	End	level	probe_count
5362355	6804250	chr1	-	120178216	120178270	full	4
5482615	6804251	chr1	-	120177258	120177332	full	4
5435005	6816537	chr1	+	41302961	41302992	full	4
5473220	6816538	chr1	+	41303198	41303285	full	4

**[GENE\_INFO]** contains gene symbols, names and other information. Examples are Gene\_Info\_Human.xls for human and Gene\_Info\_Mouse.xls for mouse. It looks like:

Probe Set Name	Identifier	EntrezGene	Name	Gene Ontology.xls	Protein Domain.xls
6747696	NM_001011873	381246	Xkr9: X Kell blood group precursor related family member 9 homolog	16020 44464 5623  16021 31224 44425	
Pathway	Chromosome	Description			
	1	Length: 32933 // Probes: 48			

**[CROSS-HYBRID\_FILE]** contains the cross-hybridization annotation of exon array probes based on our cross-hybridization analysis. Examples are Human\_CrossHyb.txt for human and Mouse\_CrossHyb.txt for mouse. It looks like:

Probeld	Probesetld	TranscriptClusterld	ProbesetLevel	MinimumProbeIntensity	MaximumProbeIntensity	MedianProbeIntensity
215959	4642578	6747200	full	-29.4633	28.3125	1.11984
4236584	4642578	6747200	full	-19.0482	165.039	2.32803
5676089	4642578	6747200	full	-21.3234	710.007	20.9295

StandardDeviation on ProbeIntensity	CorrWithTarget TcExpression	Num0bpMismatch hTranscripts	Num1bpMismatch hTranscripts	Num2bpMismatch Transcripts	Num3bpMismatch Transcripts
13.3415	0	1	0	0	1
35.8908	0	0	0	0	0
178.592	0	0	0	0	0

MaxCorrWithCrossHybTranscript	MaxCorrCrossHybld	MaxCorrCrossHyb NumberMismatches	MaxCorrCrossHyb CorePsrEvidence	MaxCorrCrossHyb RefseqEvidence
0.1823	6893918	3	1	1
0	-	0	0	0
0	-	0	0	0

#### 4. Output Files

**[OUTPUT\_FILE]:** This is the main result of MADS. This output file contains these columns: probeset\_id, transcript\_cluster\_id, chromosome, strand, start, end, level, probe\_count\_before\_cross-hybridization\_filtering, MADS p value for differential alternative splicing, probe\_count\_after\_cross-hybridization\_filtering, direction of the change in splicing index (+ for increase; - for decrease). It looks like:

4742986	6800948	chr12 - 56653227 56653304 full 4 1.40626406307e-09 4 +
4341684	6874085	chr19 - 59047244 59047515 extended 4 1.01071853522e-07 4 +
5332694	6956932	chr6 - 119919401 119919482 core 4 1.23934225725e-07 3 +
4721679	6956932	chr6 - 119919541 119919681 core 4 1.32797490654e-07 4 +
4790077	6790027	chr11 - 79059742 79059769 extended 4 1.332128746e-07 4 -

**[OUTPUT\_FILE].OUT:** This output file contains these columns: transcript\_cluster\_id, probeset\_id, probe\_id, [left side p value, right side p value], whether the probe passes extreme-value probe detection (0=extreme; 1=not extreme), average probe intensity of a probe in each of the two sample groups, average splicing index of a probe in each of the two sample groups, average transcript cluster expression index of the gene in each of the two sample groups. It looks like:

6747308	4506862	113328	0.351695259289, 0.648304740711	1	[59.279899999999998, 83.516499999999994]	[0.054871860536871341, 0.061848802764749446]	[1080.333333333333, 1350.333333333333]
6747308	4506862	1857987	0.53307541026, 0.46692458974	1	[219.5663333333332, 263.8659999999999]	[0.20323943227398952, 0.19540804739570478]	[1080.333333333333, 1350.333333333333]
6747308	4506862	3181793	0.389404888207, 0.610595111793	1	[163.7100000000001, 223.3763333333335]	[0.15153656278926259, 0.16542310540607261]	[1080.333333333333, 1350.333333333333]
6747308	4506862	6356840	0.221366463507, 0.778633536493	1	[394.1763333333333, 648.2713333333325]	[0.36486547361925337, 0.48008244877807943]	[1080.333333333333, 1350.333333333333]

---

## 5. Contact

---

The program website is <http://biogibbs.stanford.edu/~yxing/MADS/>

Correspondences regarding the MADS algorithm should be directed to Prof. Yi Xing ([yi-xing@uiowa.edu](mailto:yi-xing@uiowa.edu)) and Prof. Wing Hung Wong ([whwong@stanford.edu](mailto:whwong@stanford.edu)). Technical questions of running the MADS source code should be directed to Prof. Yi Xing ([yi-xing@uiowa.edu](mailto:yi-xing@uiowa.edu)) and Shihao Shen ([shihao-shen@uiowa.edu](mailto:shihao-shen@uiowa.edu)).