

MeV Quickstart Guide For RNA-Seq Data Getting started with RNASeq Data

Preface: This guide is an introduction to using the new RNASeq functions in MeV. The guide contains a brief tour of the new RNASeq file loader and a demonstration of a few of the new functions we have added specifically to support RNASeq data. The guide will first walk you through loading the data using the new RNA-Seq file loader. Then it will describe using an RNA-Seq-optimized module, EdgeR, to find differentially expressed genes between two groups of samples. Finally, it will demonstrate how to examine these differentially expressed genes for functional themes using the new module GOSeq.

These new options were added in MeV v4.7. If you already have MeV v4.7 installed, you can skip the **Setup** step and go directly to **Loading a Data Set**.

Setup

I. Installing MeV

- First make sure that Java is properly installed on your computer. Java v1.6 or higher for a Windows PC/Linux and v1.5 or higher for Mac OSX needs to be installed in order for MeV to work. Go to <u>http://java.com/</u> to get the latest version. Certain MeV modules also require Java 3D, which can be found here: <u>http://java.sun.com/products/java-media/3D/download.html</u>
- 2. Download the <u>RNASeq Pilot project</u>, if you have not done so already.

- 3. A screen should pop up that asks what you want to do with the files. Whether using a PC, Mac, or Linux, open the files and download them.
- 4. Once downloaded, open the folder and unzip the file.
- 5. The unzipped folder can be copied to any convenient location on the hard drive.
- 6. Open the MeV_4_7_0 folder. Double click the file called tmev.bat to run the program.

Loading a Data Set

- 1. In the Multiple Array Viewer, go to *File* \rightarrow *Load Data*.
- 2. When the window titled Expression File Loader appears, click *Select File loader* \rightarrow *RNASeq DGE Files*. The RNASeq file loader screen will appear.
- 3. Click the *Browse* button at the upper right side of the screen. In the file browser that appears, navigate to the MeV folder, then open the data/rnaseq folder. Choose the file TagSeqExample.txt. This file contains raw count data¹.

NA Seq Data Info													
	Data Type		Species	Reference Gen	ome	UCSCE	Build	Read Lengt	th				
	Count	-	Human 👻	RefSeg	-	ha19	-		Ť.				
	oconc			lititititi									
ile(Tab Delimited Multiple S	ample (*.*))												
elect data file C:\Users\e	leanora\w	orkspace	wev_rnas	eg\data\rnas	eq\T	agSegE	xamp	le.txt		Bro	wse		
elect library Si										Bro	Browse		
xpression Table													
tracking_id locus n	earest_ref	class_co	ode transcri	pt_l T1a		T1t	i.	T2	T3		N1		
Sene_00001 chr1:78931 N	R_015368	С		0	(0		2	0	0	1000		
ene_00002 chr1:14313 N	M_031921	с		20	1	8		12	5	19			
ene_00003 chr1:24951 N	M_003820	С		3	(0		2	0	0			
ene_00004 chr1:54469		-		75	1	84		241	149	271			
ene_00005 chr1:78313 N	M_004781	с		10	1	16		4	0	4			
ene_00006 chr1:78381 N	M_004781	1		129	1	126		451	223	243	_		
ene_00007 chr1:78397 N	M_004781	с		13	4	4		21	19	31			
Sene_00008 chr1:78413N	M_004781	с		0	2.4	3		0	0	0	_		
Sene_00009 chr1:80294 N	M_00112			202	1	122		256	43	287			
Sene_00010 chr1:93539 N	M_025106	i.		10	1	8		56	145	14			
Sene_00011 chr1:10075 N	M_052960	с		2	1	3		5	0	3			
ene_00012 chr1:10240 N	M_00110	с		104	(60		218	213	111			
Gene_00013 chr1:10240 N	M_00110	C		6	(6		22	13	15			
Gene_00014 chr1:10240 N	M_00110	С		0	(0		4	0	0			
	11.11		11										

The new RNASeq data loader accepts raw count data, RPKM or FPKM, mapped to either ENSEMBL IDs or RefSeq IDs.

4. Choose the appropriate parameters for each of the drop-down menus at the top of the file loader screen. For the data file we have selected, choose the *Data Type*

¹ MeV can also load RPKM data, or combined RPKM/count data. The data file isoforms.fpkm_cnt_Ref.txt is an example of this file format.

Count, the *Species* Human, the *Reference Genome* RefSeq, and the *UCSC build* hg19. Leave *Read Length* blank.

5. Click the *Load* button.

RNA-Seq Analysis

Differential Expression Detection

Begin your RNASeq analysis by testing for differential expression of all of the unique reads. To do this, we will use a module called edgeR, based on the <u>Empirical Analysis of Digital Gene Expression data in R</u> package written by Mark Robinson.

- 1. In the row of colorful buttons across the top of the MultiExperiment Viewer window, click the one labeled Statistics. Choose *Empirical Analysis of Digital Gene Expression data in R* (edgeR). An initialization dialog will appear.
- 2. Select the group membership for each of the six samples. Click "Group 1" for the first four samples, and "Group 2" for the remaining two samples.
- 3. Leave the default values for the Inference Algorithm and p-value/FDR parameters.
- 4. Click Ok. The analysis will run and display the results in the result tree, on the left of the Multiple Array Viewer window.

Multiple Array Viewer		- • • ×
Chustering	is Coppy Claudes	Visitatization
Custering Same	Istics ClisSelfication Dafa Reduction Meta Arialysis Paviidis Template Matching Tests Significance Analysis for Microarrays One-way ANOVA Two-factor ANOVA Two-factor ANOVA Nonparametric Tests Bayesian Estimation of Temporal Regulation Linear Models for Microarray Data Gene Ontology Analysis for RNA-seq Survival Analysis Global Ancova Rank Products Idutual Information Network edgeR Empirical analysis of RNA-Seq data in R DE Seq analysis of RNA-Seq data based on .ve binomial distribution DE Seq analysis of RNA-Seq data based on MA plot survival Survival	Visitalization

The edgeR module can be found in the Statistics drop-down menu.

- edgeR	Initialization			×
Experime	ent Assignments			
T1a	Group 1	Group 2	○ Excluded	^
T1b	Group 1	Group 2	○ Excluded	
T2	Group 1	Group 2	○ Excluded	
тз	Group 1	Group 2	C Excluded	
N1	Group 1 Group 2 Excluded Group 1 Group 2 Excluded			
N2	O Group 1	Group 2	○ Excluded	•
	Save s	ettings Load settings	Reset	
Paramet	ers			
	Inference algori	thm persion O Moderated tagwis	se dispersions	
	Cu	P-Value		
?	MeV* Wulti	Experiment er	cancel	ОК

The edgeR initialization dialog.

Differential Expression Results

- 1. Open up the result node labeled edgeR, and expand the nodes to find one labeled *Significant Gene List*. Click on this node to select it and display the list of genes found to be differentially expressed between the two sample groups you selected in the previous section. You can click on the links to launch a web browser displaying more information about individual genes.
- 2. Right-click on the window in a cell with no links (the *Stored Color* column is a good bet). Choose *Store entire cluster* and click Ok to label each of the genes in this window with a color. This color label will be visible anywhere a gene display is shown in MeV even in the results of other modules.

e Adjust Data Metrics Analysis Displa	v Utilities				
-3					
Custering Statistics	Classification Data Reduction	- Meta	rialysis	• Webalization	Miscelaneous
Cospinal Dats Conginal Dats C	Classification Carl Reduction Carl Reduction Core_01354(crt11664, nal_0000 Gene_01354(crt11664, nal_0000 Gene_01354(crt11664, nal_0000 Gene_12457(crt216396, nal_001 Gene_12457(crt216396, nal_001 Gene_1355(crt116345, nal_0020 Gene_1355(crt116345, nal_00200 Gene_1355(crt116345, nal_002000 Gene_1355(crt116345, nal_002000 Gene_1355(crt116345,	Intert J eff class_cc 33 c 35 c 22 c 23 c 24 c 25 c 26 c 27 c 28 c 29 c 20 c 21 i 22 c 23 c 24 c 25 c 26 c 27 c 28 c 29 c 20 c 20 c 20 c 20 c 20 c 20 c 21 c 22 c	Maryus de Transcript NA NA NA NA NA NA NA NA NA NA NA NA NA	Vitabalization (Chriel, ThT, Erect pristabalande tie664397) miclessetter, 183933172 miclessetter, 183933172 194 19409714 194097774 1940977	Miscillareous Display and the second
O General Information Script Manager History	Copy Select all rows Clear all selections Sort table in original gene order	51 c 51 c 16 c 29 c 22 c 35 c 75 c	544 544 544 544 544 544	sushi 33193637 selismo tacto 24298510 tupothetical 11817068 t440H 17934219 tupothetical 55400725 tetratricopecti 17848099	chr333101533103449 chr12429552429562 0 chr1181482118154936 chr317932082 chr255399455400532 8 chr2178479517840632
	Constant and the second s	28 c 03 c	24A 24A	plakophile 4 15053789 regulation of 15041885	1 chr2 1593134159537790 2 chr1 1503395150418710
	Broadcast Gene List to Gaggle Broadcast Selected Rows as Matrix to Gaggle Broadcast Matrix to Gaggle Broadcast Matrix to Geneme Browser	34 c 34 c 30 c 46 c	NA NA NA NA	ingine 5-m 49062408 miggen-adv 10250881 mcrobuble 39945817 protein ELYS 24701332 marten en 775521	chr3.450617549062352 chr3.450617549062352 chr3.450617549062352 chr1.25547139945476 chr1.257022247012953 chr1.2470022247012953

Results of the edgeR module, showing significantly differentially expressed genes/transcripts. Right-click to reveal a context menu with many powerful options.

Examining the differential expression list for signature themes

Now that we have a list of differentially expressed genes, we can examine it for themes. To do this, we will use the GOSeq module. This module is based on the R package <u>GOSeq</u>, by Matthew Young. It is designed to find enriched gene groups in length-biased data, such as RNASeq data. Compare it to tools like EASE for microarray data.

- 1. From the *Statistics* drop-down menu, choose the item *Gene Ontology analysis for RNA-seq.*
- 2. Leave the GOSeq parameters *Significance Level: Alpha, Number of Permutations* and *Number of Genes per Transcript Length Bin* set at their default values.
- 3. You should have a cluster pre-selected in the cluster selector dialog. If you have more than one cluster available in this dialog, choose the one you want to examine for geneset enrichment.
- 4. Choose *Download from GeneSigDb* from the drop-down menu. Click the *Download* button.
- 5. Check that the *Choose Annotation Type* drop-down menu is set at *GENE_SYMBOL*.
- 6. Leave the *File Location* field blank.
- 7. Click Ok. GOSeq will run.

DSEQ Paran	meters						
			Significance Leve	st Alpha = 05	i		
			agained to re	ne mprio - <u>105</u>			
			Number of Perm	nutations: 1000]		
		Number	of Gapos par Transcript I	anoth Bier 20	1		
		Pie	tor delives per transcript of	Englin black 20	1		
luster Grap	h		ase select a cluster of dimension	nany expressed genes			
ene Cluste	rs						
	Source	Factor	Cluster Node	Cluster Label	Remarks	Size	Color
1	Algorithm	edgeR (1)	Significant Gene List			485	
1	Algorithm	edgeR (1)	Significant Gene List			485	
ne Set File	Algorithm	edgeR (1)	Significant Gene List	nn Geour SigDel 🛛 🔻		485	Downioa
ne Set File Choose	Algorithm File Loaded e Annotation Type:	edgeR (1)	Significant Gene List	HE Generation +		485	Downloa
ne Set File	Algorithm File Loaded e Annotation Type: File Location: C1Us	ersieleanoral.mev/reposit	Significant Gene List	ITS Come SigDE + NE_SYMBOL • duster gui impl gsea Ger	reSigDbGeneSetsige	485	Downsiga

The GOSeq initialization dialog.

Signature theme results

In the Result Tree, you will see a new result node named GOSEQ.

- 1. Open this node and select the node labeled *Results Table*. This table contains the complete list of genelists downloaded from the GeneSigDb database, as well as a rating for each list as to wether the contents of that list is enriched in the selected group of differentiated genes used to run GOSeq.
- 2. Double-click on the header labeled *p-value* to sort the list. Those gene lists with low p-values, like Human StemCell_Brendel05_21genes, listed here, are enriched in the set of differentially expressed genes we found in our previous edgeR analysis. You can explore this gene list by going to the GeneSigDb website.

	-		-	-		12		100 C	T
Custering Statistics	Classification	Data Reduction	•	Meta Artalys	is 🔻	Visualization	- Mi	cellaneous	ł
	1	Gene List			Gene Court	Significant Gen.	Expected St	g g-value (perm	iii.
	Human StemCell_E	Brendel05_21genes			4.0	1.0	0.01844728	0.021	
Original Data	Human Breast_Sor	ng06_61genes			0.0	1.0	0.027549945	5 0.029	
Cluster Manager	Human Breast_He	ndricks04_54genes_BRG1_	sowm		8.0	1.0	0.036809664	3 0.031	
Sample Clusters	Human Prostate_H	lendriksen06_60genes			9.0	1.0	0.04132855	7 0.033	
HB Cana Chusters	Human Prostate_S	hepherd08_50genes			10.0	1.0	0.04527040	3 0.042	
HII Office Closelles	Human Breast_Lee	H06_71genes_up_VWD3			10.0	1.0	0.04592052	5 0.045	
Analysis Results	Human Lung_Cold	ren06_110genes			16.0	1.0	0.07361015-	4 0.082	
Data Source Selection	Human Breast_Ch	ango4_rrzgenes_WoundHe	anng_CS	sk_genes	122.0	2.0	2.0 0.56317574		
edpeR (1)	Human Breast_De	smedius_229genes_AURKA	Module	all a second alla	37.0	1.0	0.1/07577	0.157	
Expression Images	Pruman Bysast_Var	scent yndenur 259genes-h	poss_as	gogenästs	38.0	1.0	0.17474295	0.174	
Time Score Canad	Human Breast_Int	pengenuo_onzgenes_ruro			04.0	1.0	0.2920401	0.227	
a organicani ornito	Human Breast Lat	All Transet			11.0	1.0	0.32/9//15	0.3	
Non-significant Genes	Human Coles, Ion	top_17 spenes			103.0	1.0	0.4/538305	0.379	
P III Table Views	Human Draast Ch	sating asymits			110.0	1.0	0.2440926	0.413	
Gene List	Human Dreast_Crit	arare-sauneuro_1505genes			161.0	1.0	0.7419625	0.598	•
HB Similicast Cana List	Human Ovarian_Baranova06_907genes				199.0	1.0	0 87634447	0.640	
THE New Presidents Constitut	Human Breast Kite	ike07 3712nenes Rasal m	abs.		495.0	1.0	2 2896047	0.89	
Effluton-Sidnarcaus Ownis Cras	Human Breast Far	mer05 3198nenes basal a	annon	luminal	519.0	1.0	2 3938997	0.911	
General Information	Human Breast, Reval05, 3373genesSuppTable2SuppTable2			tie?	6310	1.0	2.4*02487	0.921	
GOSEQ (2)	Human Breast, Baloph05, 90genes			10.0	0.0	0.046327226	10		
Expression Images	Human Breast, Wa	no05 230genes down BT-	AF57cel	ts.	49.0	0.0	0.22593704	1.0	
Gifferentially Expressed Cluster	Human Endometria	M Salvesen09 11cenes			1.0	0.0	0.004591594	4 1.0	
Mon Differentially Expression Churcher	Human Colon_Arar	ngo04 254genes			30.0	0.0	0.13820705	1.0	
The second secon	Human Breast, Bal	ogh06_23genes_known			2.0	0.0	0.00918403	1.0	
P III Table views	Human Endometria	al_Matsushima-Nishiu01_69	penes		4.0	0.0	0.018358661	1 1.0	
Differentially Expressed Cluster	Human Colon_Aller	nD8_15genes			2.0	0.0	0.009384371	3 1.0	
Non-Differentially Expressed Cluster	Human Breast_Vin	cent-Salomon08_502genes			157.0	0:0	0.72381	1.0	
EIII Results Table	Human Liver_LI02_	10genes			3.0	0.0	0.01377526	7 1.0	
Prohability Weighting Evention	Human Liver_LX02_	_38genes			7.0	0.0	0.03252816	1.0	
C Producing weighting Punction	Human Colon_Aller	nD8_25genes			5.0	0.0	8 023070076	3 1.0	
- U General Informatión	Human Prostate_C	handran07_100genes			3.0	0.0	0.01421578	5 1.0	
Script Manager	Human Bladder_O	sman06_346genes			65.0	0.0	0.30172363	1.0	
History	Human Colon_Del	laRagione01_21genes			7.0	0.0	0.032405667	1.0	
	Pruman Prostate_C	mandran07_154genes			23.0	0.0	0 1053471	1.0	
	Human Prostate_C	handran07_208genes			22.0	0.0	0.10174374	1.0	
	Human Bladder_Or	sman06_346genes			65.0	0.0	0.30172363	1.0	
	Mouse threast_Why	phos_339genes			1.0	0.0	0.004591591	1 1.0	
	Human Colon_dP	etroos_100genes			18.0	0.0	0.08312415	1.0	

Gene signatures, published in GeneSigDb, with enrichment in the list of selected genes. Future plans include adding links from this display directly to the gene signature web page, where the list of genes in the signature and the source publication can be found.

From here, you can continue examining gene signatures of interest by searching the GeneSigDb website, or continue on with another analysis by simply selecting it from one of the drop-down menus. For this pilot, most of the standard MeV modules are available to use. A few of them, like the EASE and GSEA modules, require specific annotation files that are currently only available for DNA micoarray data. Part of the full RNASeq implementation project will be to adapt MEV to fully support RNASeq analysis in all modules. However, that support is not yet available.

Additional Resources

For more information on the different modules, consult the MeV Manual, found in the "documentation" folder, or go to http://mev.tm4.org.